

UE3

# Journal de bord City'nder

Le groupe n° 3  
Pour le 02/02/24

<b>I JEU(X) DE DONNÉES</b>		<b>3</b>
1	<b>LE JEU DE DONNÉES INITIAL</b>	<b>4</b>
1.1	Choix et objectifs	4
2	<b>LES JEUX DE DONNÉES COMPLÉMENTAIRES</b>	<b>7</b>
2.1	Les enrichissements	7
<b>II TRAITEMENT DES DONNÉES</b>		<b>12</b>
3	<b>PRÉAMBULE</b>	<b>13</b>
3.1	Modèle de base de données relationnelle	13
4	<b>LES REQUÊTES SPARQL</b>	<b>14</b>
4.1	Explication des requêtes	14
5	<b>LE FLUX DATAÏKU</b>	<b>17</b>
5.1	Une vision d'ensemble	17
5.2	Traitement sur le dataset initial	17
5.3	Traitement sur les enrichissements	19
<b>III DATAVISUALISATIONS</b>		<b>26</b>
6	<b>PRÉAMBULE</b>	<b>27</b>
6.1	Avant-propos	27
7	<b>LES DATAVISUALISATIONS</b>	<b>28</b>
7.1	Illustration et analyse	28
8	<b>CONCLUSION</b>	<b>42</b>
8.1	A retenir pour l'avenir	44

# JEU(X) DE DONNÉES

# 1. LE JEU DE DONNÉES INITIAL

## 1.1. CHOIX ET OBJECTIFS

Dans le cadre du Master 2 Technologies numériques appliquées à l'histoire à l'Ecole nationale des chartes, le groupe n°3, constitué de Gilmar Ballestrim, Marina Hervieu, Anna Le Duff et Sarah Marcq, a pour objectif de réaliser une application web nommée **City'nder**. Pour réaliser ce travail, il convient de collecter des données, les nettoyer, les enrichir, les analyser, en proposer des visualisations pour finalement les utiliser en constituant une base de données, nécessaire au développement de l'application web. Le dataset initial a été choisi par l'ensemble du groupe. Puis, chaque membre de l'équipe s'est occupé de choisir un ou plusieurs datasets complémentaires, de le traiter puis d'effectuer une datavisualisation à partir du dataset final. Ensemble, nous avons travaillé au traitement global et les tâches ont été plus ou moins réparties de façon équitable. Surtout, chacun a apporté son expertise et sa créativité pour aboutir à traitement de données et un projet d'application pertinents. City'nder propose à l'utilisateur des communes **où vivre en fonction de critères** qu'il a lui-même complétés au fur-et-à-mesure de sa navigation sur l'application. Le premier indicateur est le **prix du loyer** au m<sup>2</sup> (maison et appartement) qui est représenté par le dataset initial et les **autres critères** (culture, sport, nature, commerces, tourisme) font l'objet d'enrichissements. L'ensemble est décrit en détail dans les parties suivantes.

### 1.1.1. Une description du jeu de données initial ...

Issu du site **Datagouv.fr**, le dataset initial répertorie les loyers (appartements et maisons) en France (hors Mayotte) jusqu'en novembre 2022 à partir des annonces des siteweb Le Boncoin et Seloger.com . Les loyers sont présentés par communes (ou par arrondissements pour les grandes communes) avec leur code Insee, qui représente le principal avantage pour notre travail. En effet, l'ensemble des enrichissements pourra se joindre grâce à cette donnée élémentaire.

Dans le dataset, en plus du code Insee et des indicateurs de loyers, nous disposons de données géographiques : le libellé géographique des communes et EPCI, leurs coordonnées géographiques (longitude, latitude), le numéro des départements et des régions. Du côté des loyers, nous avons aussi les indicateurs les plus élevés et les plus bas. Nous avons aussi accès au « Coefficient de détermination » («R2-adj»), qui sert à déterminer l'indicateur du prix du loyer : plus il est élevé plus l'indicateur de loyer est proche des loyers observés dans les annonces. Le nombre d'observations par commune et par maille permet aussi de déterminer la fiabilité de l'indicateur, s'il est inférieur à 30 il est moins fiable, et inversement.

Il faut noter que le présent dataset, que nous appelons initial pour notre travail, s'inscrit dans un corpus plus large de dataset :

1. **Indicateurs de loyer appartement** ;
2. Indicateurs de loyer appartement de 1 ou 2 pièces, ;
3. Indicateurs de loyer appartement de 3 pièces et plus ;
4. **Indicateurs de loyer maison**.

Nous avons choisi de ne sélectionner que les deux dataset qui étaient les plus généraux soit ceux soulignés en **gras** ci-dessus.

### 1.1.2. ... aux avantages ...

Le choix s'est porté sur ce dataset car il présente plusieurs avantages.

- La **propreté**. Rares sont les cases vides ou non complétées, les accents transformés en caractère illisible ou encore des séparateurs (virgules, points).
- La **complétude**. Les données sont complètes car des informations supplémentaires sont ajoutées à l'information principale. Au prix du loyer par commune (qui est l'information centrale et le résultat d'un calcul), le dataset répertorie aussi la région, le département, les coordonnées géographiques de la commune, et bien sûr le nom de la commune et surtout, son code Insee. Les données sont ainsi compréhensibles car explicites dans leur dénomination.
- La **documentation**. Le site de référencement des dataset, fournit une documentation riche (un FAQ, une note méthodologique détaillée, un guide d'utilisation des données et un dictionnaire des variables) permettant une bonne compréhension des méthodes de calculs utilisées pour aboutir à ces données.
- Le **principal avantage** est le **code Insee**. Cet identifiant unique de chaque commune en France est un repère fondamental, très utilisé en statistique géographique, et potentiellement dans d'autres dataset répertoriés par communes. Ce point, très important, constitue le principal atout de notre travail car il paraît assez aisé de joindre plusieurs sets de données à notre tableau initial.

Finalement, la simplicité, la richesse de la nomenclature du dataset initial offrent un bouquet de possibilités et beaucoup d'inspiration pour le projet final.

### 1.1.3. ... et aux risques ...

Cependant, nous avons rencontré des problèmes lors de l'analyse plus en profondeur du dataset. Ils sont majoritairement liés aux sujet de celui-ci alors source de risques dont nous avons conscience.

- Le dataset initial relève de l'**économie, du logement, des statistiques** : autant de domaines très éloignés de notre domaine d'expertise (histoire, histoire de l'art, sciences humaines). Ces notions sont complexes à comprendre en profondeur, surtout si, comme nous, on ne dispose pas de connaissances préalables sur le sujet. Nous avons donc pris du temps à comprendre les composantes du dataset (notion de maille, idzone et les indicateurs plus ou moins proche de la maille), ainsi que les méthodes utilisées pour obtenir les données (machine-learning). Même si nous avons fait l'effort d'approfondir ces sujets, il y a bien des notions qui restent floues et hors de notre portée en tant qu'étudiants en Master TNAH.
- Un deuxième risque est d'apporter une **mauvaise analyse et une surinterprétation** de nos données, ce qui pourrait nous mener à des résultats erronés ou biaisés et de voir des corrélations qui n'existent pas.
- Un troisième risque concerne le **périmètre temporel et spatial** de notre dataset initial : la géographie des communes est celle de novembre 2022. Ce n'est pas d'actualité avec la géographie du dernier trimestre 2023 au premier trimestre 2024 - bornes temporelles de nos phases de travail. Ce risque est aussi à prendre en compte lors des enrichissements des données au dataset initial.

Comme nous avons conscience de ces risques et de notre manque de connaissance sur le sujet, nous avons décidé de faire très attention aux analyses que nous allons proposer et de garder en tête la **marge d'erreur** : moins les résultats sont fiables et plus la probabilité

qu'ils soient écartés de la réalité est importante.

#### 1.1.4. ... déterminant pour notre problématique.

A la lecture et à l'étude du dataset initial, nous avons défini un **angle d'attaque** pour l'ensemble de notre travail : la **commune est la dimension générale** que l'on veut aborder alors sous différents facettes grâce aux enrichissements présentés dans la sous-partie suivante.

## 2. LES JEUX DE DONNÉES COMPLÉMENTAIRES

### 2.1. LES ENRICHISSEMENTS

Pour répondre à l'objectif de notre application, nous avons besoin d'enrichissements variés qui apportent des paramètres pertinents dans la recherche de la commune idéale où l'utilisateur voudrait se loger.

Nous nous sommes accordés pour que chacun enrichisse le dataset initial d'au moins un dataset complémentaire. Ces datasets proviennent de différentes sources : *Datagouv*, *Insee*, *Observatoire des territoires* ... Ci-suit sont présentés les enrichissements réussis et aussi quelques échecs et difficultés rencontrés.

#### 2.1.1. Des réussites ...

##### Sport

La présence d'**équipements sportifs** dans une commune est un **facteur déterminant** pour un profil utilisateur entièrement orienté sport ou pour un profil dont le sport est un des critères de sélection. C'est pourquoi, il nous semble pertinent d'ajouter une liste d'équipements sportifs (stades, salles de sports, piscine, etc...) à nos données initiales.

Tout d'abord, nous avons trouvé un dataset sur **data.gouv.fr**, qui est un recensement des équipements sportifs, des espaces et des sites de pratiques. Intéressant de prime abord, il n'a finalement pas été retenu car il sépare distinctement, en deux dataset, les équipements des installations sportives. Intégrer ce double jeu de données aurait rendu le travail plus compliqué que nécessaire. De plus, nous nous sommes aperçus qu'il répertorie des installations qui ne correspondent pas *stricto sensu* à notre critère sportif : aux gymnases municipaux sont accolés les écoles primaires voire des centres pénitentiaires.

Finalement, nous avons trouvé un dataset qui correspond davantage à nos critères : du nom de **Data ES**, il référence une **base de données** extrêmement complète sur les équipements sportifs et des lieux de pratiques en France. Crée par le ministère des Sports, des Jeux Olympiques et Paralympiques, elle est mise à jour quotidiennement (en vue des JO en 2024)<sup>1</sup>. Dans ce dataset, nous disposons d'un grand nombre de données et nous avons rapidement déterminé qu'il y a avait tout ce qu'il nous fallait pour effectuer notre analyse.

Chaque installation sportive est nommée, numérotée et située par commune à laquelle sont référencés le département et la région. Le **code Insee** est répertorié pour chacun des équipements, ce qui nous permet de le relier facilement avec les données de notre dataset initial ainsi qu'avec nos autres enrichissements. Nous avons aussi accès à des informations plus générales sur le type de l'équipement, la famille de l'équipement, et une catégorie « Atlas » est proposée. Ces données nous ont semblé essentielles pour l'analyse du dataset car elles permettent de manipuler les équipements sportifs à partir de leur nature, ce qui aurait été beaucoup plus difficile à faire si nous n'avions que le nom de l'établissement, rarement très transparent.

1. Cette base de données étant mise à jour quotidiennement, la version dont nous disposons et que nous avons traité n'est pas la même que celle qui est présente en ligne aujourd'hui.

## Culture

En complément, nous pensons qu'il est opportun d'ajouter un élément culturel à notre jeu de données initial – ne serait-ce que pour se rapprocher de notre domaine de connaissances.

C'est ainsi que notre jeu de données sur les **infrastructures culturelles** est issu de **Basilic**, la base des lieux et équipements culturels de la France entière. Conçue dans le cadre de la conception de l'Atlas régional de la Culture et administrée par le département des études, de la prospective, des statistiques et de la documentation du ministère de la Culture, elle est réalisée par agrégation de différentes sources élaborées par plusieurs directions du ministère – c'est pourquoi quelques erreurs et doublons subsistent.

Toutefois, cette **base géocodée** offre de nombreux avantages : un **volume considérable de données, une mise à jour des données récente, et, une nomenclature claire**. En effet, de par un nombre conséquent de contributeurs, 73 234 équipements culturels sont enregistrés depuis le dernier traitement effectué en août 2023. Les labels et appellations du patrimoine, du livre, du cinéma et de la création, classent les structures culturelles offrant ainsi plus de 40 catégories différentes. Surtout, l'ensemble est organisé par variables au libellé explicite et au contenu similaire à notre jeu de données initial. Le **code Insee** par exemple, clé centrale de tous nos enrichissements, est référencé pour chaque structure. Le libellé géographique (le nom de la commune) est également présent et est important à connaître car la géographie française s'actualise d'une année à l'autre en raison de la disparition de certaines communes alors regroupées en EPCI.

Par conséquent, ces éléments viennent justifier notre choix : nous pensons que l'**offre culturelle** d'une commune, d'un département ou d'une région est un **indicateur non négligeable** pour l'utilisateur de notre application.

## Commerce

Ensuite, il nous paraît notable qu'un utilisateur de notre application puisse savoir si la commune qui correspond à ses critères est pourvue de **commerces**.

Pour ce faire, un enrichissement sur les commerces vient compléter notre jeu de données initial. La **base permanente des équipements** (BPE) de l'Insee, offre pléthore d'équipements par thématiques (sports, loisirs, cultures, commerces, services aux particuliers, action sociale, services de santé, tourisme et transports, établissements d'enseignements, etc.) et nous pensons que celle qui dénombre les équipements et les services dans le domaine du commerce est pertinente. De plus, cette source statistique est déclinée en deux jeux de données aux échelles différentes : l'une infracommunale et l'autre **communale** - évidemment sélectionnée car la commune est l'élément primaire de notre dataset initial.

Elle liste des commerces de nature très variée : des hypermarchés, supermarchés, grande surface de bricolage, supérette, épicerie, boulangerie, boucherie/charcuterie, produits surgelés, poissonnerie, des magasins en tout genre (de chaussures, équipements du foyer, électroménagers et matériel audio-vidéo, de meuble, de sports et loisirs, de bricolage), les quincailleries, parfumerie et esthétique, fleuriste, stations-services, etc. Cet enrichissement risque d'être très volumineux, il conviendra de le traiter avec précaution.

Enfin, cette base a bien évolué car éditée en 2016, elle a été mise à jour en juillet 2022 - ce qui correspond (à quelques mois près) à la géographie des communes en France de notre dataset initial (novembre 2022).

## Nature

En ce qui concerne l'**environnement géographique**, notre but est de déterminer si une commune présente des conditions environnementales particulières. Nous avons choisi de nous concentrer sur les **zones de montagne, les zones littorales et les parcs naturels régionaux et nationaux**. Nous souhaitons mettre en avant les littoraux et les massifs parce qu'il s'agit de données explicites et pertinentes pour un utilisateur. Nous avons également sélectionné les parcs naturels régionaux et les parcs nationaux car ce sont des territoires dont l'environnement est davantage préservé. Leur présence sous-entend une certaine richesse en termes de **patrimoine naturel** et culturel et une plus grande protection de la biodiversité. Nous avons ainsi mis l'accent sur l'idée de protection. Les aménagements sont contrôlés dans l'ensemble de ces zones.

Nos datasets concernant l'environnement géographique ne sont en rien exhaustifs. Nous aurions par exemple aimé ajouter les réserves naturelles et les sites Natura 2000. Nos choix ont été effectués en fonction des datasets disponibles, de leur qualité et des facilités de jointure avec le dataset de départ sur les loyers. Les datasets choisis ont également tous été mis à jour récemment (en 2022 ou 2023).

Les datasets choisis sont les suivants :

- **Les communes des Parcs nationaux (PN)**

Ce dataset issu de l'*Observatoire des territoires* indique si une commune française appartient à un parc national ainsi que le nom de celui-ci.

- **Les communes des Parcs naturels régionaux (PNR)**

Ce fichier provient de la même source et est construit de la même façon que celui sur les parcs nationaux.

- **Les communes concernées par la Loi Littoral**

Provenant de *data.gouv.fr* et produit par le *Ministère de la Cohésion des territoires*, ce dataset liste les communes soumises à la loi littoral et si ce classement est issu d'une proximité avec la mer, un estuaire ou un lac. En effet, promulguée en 1986, cette loi encadre l'aménagement côtier. Elle concerne des communes "riveraines des mers et océans, des étangs salés, des plans d'eau intérieurs d'une superficie supérieure à 1 000 hectares" et "riveraines des estuaires et des deltas lorsqu'elles sont situées en aval de la limite de salure des eaux et participent aux équilibres économiques et écologiques littoraux".<sup>2</sup>

- **Les communes concernées par la Loi Montagne**

Ce dataset a la même source que celui de la Loi Littoral et est construit de la même façon. La Loi Montagne a été promulguée en 1985 puis complétée en 2006. Elle concerne les communes situées dans les zones de massifs.

- **Les périmètres de massifs**

Le dataset sur les périmètres de massifs vient compléter celui de la Loi Montagne. Il est issu de *data.gouv.fr* et produit par l'*Agence nationale de la cohésion des territoires*. Les chaînes de montagne concernées sont celles des Vosges, du Jura, des Alpes, le Massif central, les Pyrénées, ainsi que les zones de massif de Corse, de Martinique, de Guadeloupe et de la Réunion. L'intérêt est de pouvoir récupérer un libellé pour le massif et de

2. *Ministère de la transition écologique et de la cohésion des territoires*, «Loi relative à l'aménagement, la protection et la mise en valeur du littoral», url : <https://www.ecologie.gouv.fr/loi-relative-lamenagement-protection-et-mise-en-valeur-du-littoral>

pouvoir mieux quadriller les communes qui ont un intérêt naturel spécifique.

## Tourisme

Lorsqu'un utilisateur recherche un logement, on peut émettre l'hypothèse qu'il ne souhaite pas vivre dans une zone de tourisme - surtout si les logements ne sont en réalité que des hébergements touristiques. C'est pourquoi, il nous paraît intéressant de le lui faire signaler en intégrant à notre jeu de données initial, un ensemble consacré aux **hébergements touristiques**. Ce jeu de données datant de 2023 présente les indicateurs clés relatifs aux capacités d'hébergement touristique des communes, englobant les hôtels, campings et autres formes d'hébergement collectif en France *hors Mayotte* en géographie 2022. Il inclut également des données sur la répartition des emplacements de campings, distinguant entre ceux loués à l'année et ceux proposés à une clientèle de passage. En ce qui concerne Dataiku, cette base de données revêt une importance particulière, car elle pourrait fournir d'emblée des idées sur la corrélation entre les loyers et l'activité touristique.

Ce jeu de données a été généré par l'Insee et est alimenté par l'Insee ainsi que par des partenaires territoriaux, des sources que nous considérons fiable. Par ailleurs, la cohérence géographique avec notre propre jeu de données est assurée, puisque les deux ont été élaborés en se basant sur la géographie de l'année 2022, malgré le fait que la diffusion du jeu de données ait été faite en 2023.

La présence de données répertoriées par code Insee a confirmé notre choix, car elle permet un lien facile avec notre dataset initial. De plus la base est très bien documentée, avec les définition des variables, des informations sur les indicateurs et la qualité des données, ce qui a permis une bonne compréhension des données.

Par ailleurs, nous estimons que ce jeu de données revêt une importance particulière pour les utilisateurs de notre application. Cette focalisation spécifique sur les tarifs de location dans une zone touristique, que ce soit pour une résidence à long terme près de la mer, une location temporaire de quelques mois en raison d'une mission, ou en lien avec des considérations professionnelles liées à la proximité du lieu de travail (que l'utilisateur soit travailleur ou propriétaire), renforce cette pertinence.

### 2.1.2. ... aux échecs

#### Médical

Pour compléter la notion de services, nous pensons que connaître l'**offre médicale** d'une commune est important. Cependant, **aucun enrichissement** consulté n'était **convaincant** : soit vieilli par leur date de mise à jour, soit **incomplet** par l'absence de liste exhaustive sur les médecins généralistes et l'adresse de leur cabinet. Il aurait été envisageable de scrapper le site Doctolib et d'indiquer en attribut l'adresse du médecin et la commune associée. Cependant, nous ne pensons pas que cela soit une pratique à utiliser dans le cadre de notre travail universitaire. De plus, nous sommes convaincus que nous avions déjà suffisemment de jeux de données à traiter.

#### Pollution lumineuse

Pour finaliser notre travail, nous voulions ajouter un **indicateur amusant** pour l'utilisateur. L'idée était de proposer le critère "**Voir les étoiles en se couchant le soir**". La

source initiale provient d'une **association d'astronomie** qui cartographie la pollution lumineuse en cherchant à lutter contre celle-ci. Mais ce jeu de données ne s'avérait finalement **pas libre de droit**. Quand bien même, il l'aurait été, les données demandent un certain temps de compréhension et le traitement aurait été très difficile pour notre niveau de compétence car la pollution lumineuse n'est pas enregistrée selon les coordonnées précises d'une commune mais plutôt d'une zone déterminée par plusieurs points. Enfin, même si nous aurions réussi cet exploit, le temps d'intégration et de chargement à l'application finale aurait été trop volumineux<sup>3</sup>.

---

3. La réalisation d'une carte a demandé plus de 11 heures de chargement juste pour l'ouverture et la sauvegarde du fichier de la carte soit 1.98To - des mesures hors de notre portée dans le cadre de ce travail.

# **TRAITEMENT DES DONNÉES**

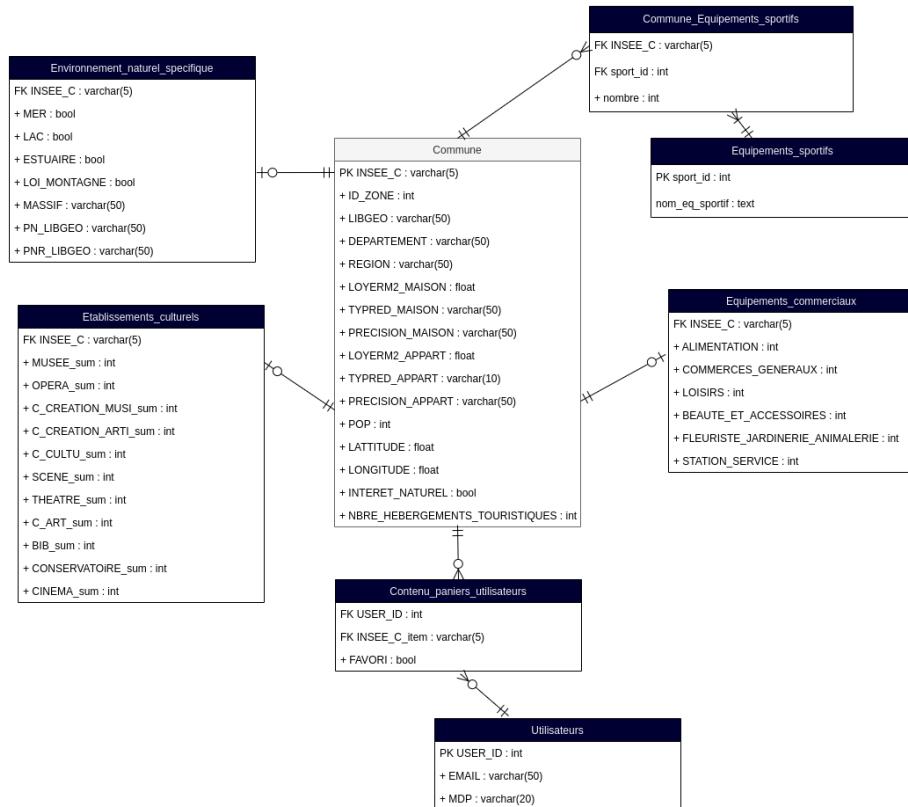
### 3. PRÉAMBULE

#### 3.1. MODÈLE DE BASE DE DONNÉES RELATIONNELLE

Notre traitement de données a **deux objectifs** : **fournir une table avec des informations pertinentes et comparables pour les datavisualisations**, que nous pourrions facilement importer dans *Tableau*; et dégager et **organiser les données** qui seraient contenues dans notre future **base de données**.

Nous souhaitons éviter d'avoir à effectuer de nouveaux traitements dans les csv récupérés pour construire la base. Nous voulions en effet pouvoir récupérer les datasets intermédiaires avant la jointure finale pour créer les tables secondaires. Nous voulions aussi que le dataset final du flux extrait de *Dataiku* forme la table principale sans avoir à le nettoyer. Des colonnes inutiles à l'application seront simplement supprimées.

La base de données se présente ainsi :



Dans la partie précédente, nous avons insisté sur le **code Insee** qui est la clé qui relie l'ensemble de nos tables. La table principale de la base de données diffère légèrement de celle dont nous avons besoin pour les datavisualisations. Elle nécessite en effet une suppression de ses éléments requêtables par un simple «`.count()`» dans Flask-SQLAlchemy. Une table de relation sera créée par Maxime Challon pour les équipements sportifs à partir d'un csv que nous lui transmettrons.

# 4. LES REQUÊTES SPARQL

## 4.1. EXPLICATION DES REQUÊTES

Plusieurs requêtes SPARQL ont été effectuées. Le dataset initial était assez complet. Le choix des informations requêtées a été pensé **en fonction des besoins des datavisualisations**. Nous avons décidé de nous tourner vers *Wikidata*. En effet, le SPARQL endpoint de wikidata est très intuitif, les données sont variées et assez précises.

Variables sélectionnées

Information sélectionnée	Nom de la variable	Propriété liée	Objectif
Code Insee	?codeINSEE	wdt:P374	jointure avec le dataset initial
Coordonnées géographiques	?loc	wdt:P625	réaliser une carte
Population	?population	wdt:P1082	répondre aux besoins de l'application et des datavisualisations
Superficie	?superficie	wdt:P2046	répondre aux besoins de l'application et être en mesure d'intégrer les densités de population dans nos analyses sur Tableau
Date de modification	?modification-Date	schema:datemodified	répondre à un problème de doublons détaillé ci-dessous
Nom de la commune	?cityLabel	ici obtenu automatiquement grâce au suffixe "Label" ajouté à city dans les variables sélectionnées	mieux se repérer dans les données.

```

1 SELECT ?cityLabel ?population ?loc ?superficie ?codeinsee
2 WHERE {
3   ?city wdt:P31 wd:Q484170 ;
4     wdt:P1082 ?population ;
5     wdt:P625 ?loc ;
6     wdt:P2046 ?superficie ;
7     wdt:P374 ?codeinsee .
8
9
10 SERVICE wikibase:label {
11   bd:serviceParam wikibase:language "fr" .
12 }
13 }
```

Première requête SPARQL effectuée

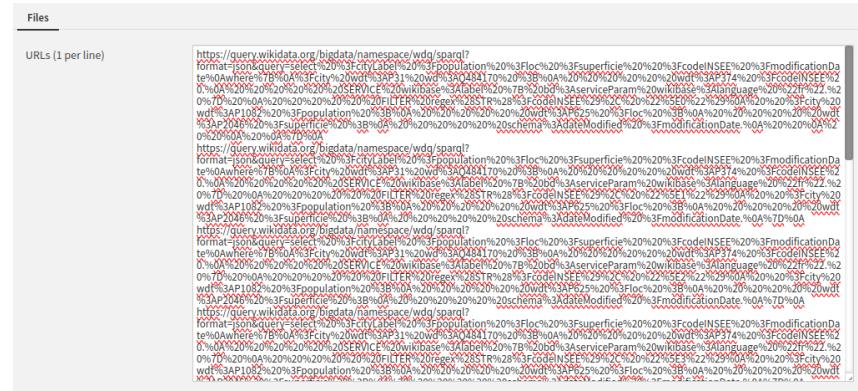
La première requête effectuée concerne les **communes françaises** (Q484170). Écrite telle que ci-dessus, elle produit une quarantaine de milliers de résultats. Étant trop lourde et par conséquent lente à s'effectuer, elle posait problème dans Dataïku au moment de l'import. Le choix a été fait de **diviser la requête**. Pour cela, nous avons d'abord requêté les communes dont le code Insee commençait par 0, puis par 1 et ainsi de suite jusqu'à 9 grâce à un **filtre basé sur les expressions régulières**. Cela a été assez rapide dans la mesure où il suffisait de copier-coller un premier url et de changer un seul caractère.

Un second problème s'est posé par la suite. Il s'agit d'un problème de **doublons**. Des informations ont été modifiées au fil du temps pour certaines communes, or l'interrogation de Wikidata ne renvoie pas seulement le dernier état de l'information. Nous avons gardé la requête telle quelle en ajoutant une colonne pour la **date de modification**, afin de supprimer les doublons les plus anciens lors du traitement sur Dataïku.

```

1 select ?cityLabel ?population ?loc ?superficie ?codeINSEE ?modificationDate
2 where{
3   ?city wdt:P31 wd:Q484170 ;
4     wdt:P374 ?codeINSEE .
5     FILTER regex(STR(?codeINSEE), "^1")
6   ?city wdt:P1082 ?population ;
7     wdt:P625 ?loc ;
8     wdt:P2046 ?superficie ;
9     schema:dateModified ?modificationDate.
10
11 SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }
12 }
```

Exemple de requête effectuée pour les communes



## Interface pour les ajouts des URL dans Dataiku

Par ailleurs, les requêtes sur les communes françaises laissent de côté les **arrondissements**, nécessitant ainsi une interrogation distincte sur Wikidata. Pour cette dernière étape, nous avons choisi de sélectionner les localisations administratives (P131) de Paris (Q90), Lyon (Q456) et Marseille (Q23482). Certains arrondissements ne sont pas directement liés à leur ville. Il faut alors passer par d'autres localisations administratives, d'où la nécessité de sélectionner les objets liés à la ville par un ou plusieurs chemins contenant la propriété *P131* (wdt:P131/P131\*). L'interrogation se fait de manière récursive sur les localisations administratives de villes et les localisations administratives de ces localisations administratives. Parmi elles on ne sélectionne que les items qui sont des arrondissements municipaux (wd:Q702842). Le reste du contenu de la requête est le même que celui de la requête concernant les communes pour pouvoir s'aligner avec elle.

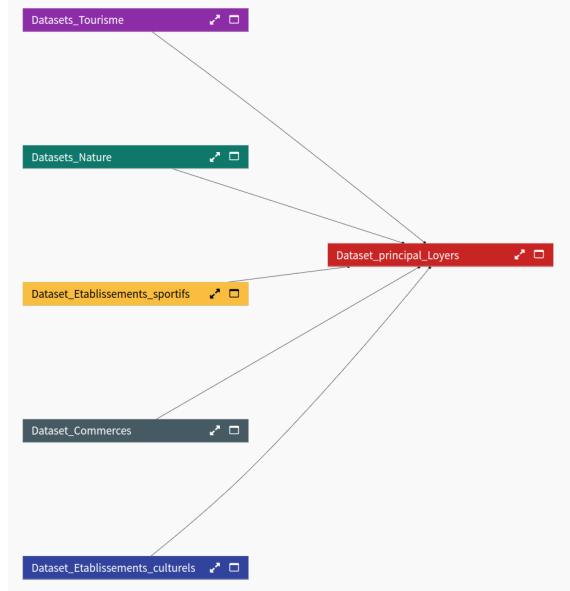
```
1 select ?cityLabel ?population ?loc ?superficie ?codeINSEE ?modificationDate
2 where{
3 ?city wdt:P131/wdt:P131* wd:Q90}
4 UNION
5 ?city wdt:P131/wdt:P131* wd:Q456}
6 UNION
7 ?city wdt:P131/wdt:P131* wd:Q23482}
8 ?city wdt:P31 wd:Q702842 ;
9     wdt:P374 ?codeINSEE ;
10    wdt:P1082 ?population ;
11    wdt:P625 ?loc ;
12    wdt:P2046 ?superficie ;
13    schema:dateModified ?modificationDate .
14
15 SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }
16 }
```

## Requête effectuée pour les arrondissements

# 5. LE FLUX DATAÏKU

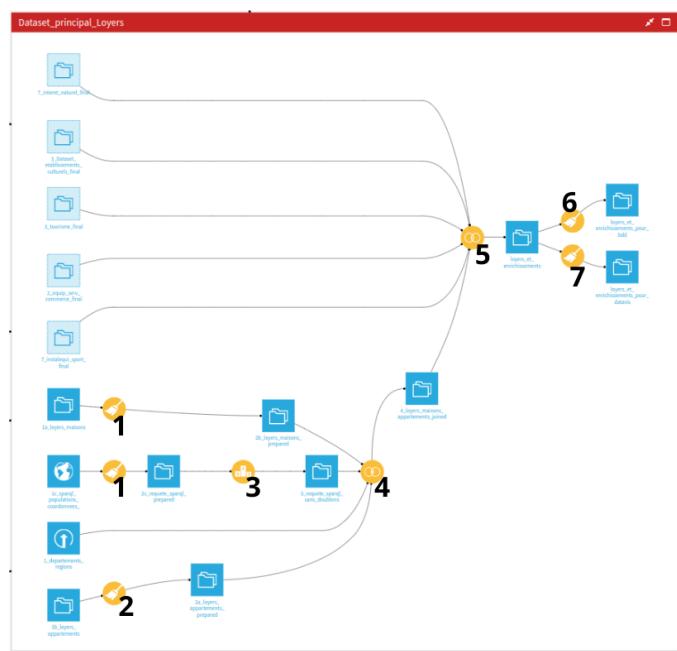
## 5.1. UNE VISION D'ENSEMBLE

Pour présenter le flux Dataiku, nous détaillons l'ensemble des opérations effectuées sur les jeux de données, depuis le jeu initial et des enrichissements jusqu'au jeu de données final.



Vision globale du flux

## 5.2. TRAITEMENT SUR LE DATASET INITIAL



Section du flux consacrée au dataset initial et final sur les loyers

0. **Datasets** . Nous avons choisi de travailler sur le prix du mètre carré pour les appartements et pour les maisons. Nous avons donc importé ces deux fichiers, ainsi que les résultats de la requête SPARQL détaillée précédemment, cette dernière ayant été réalisée directement dans Dataiku. Un dataset provenant de [data.gouv.fr](http://data.gouv.fr) contenant **les noms des départements et des régions françaises** a également été importé afin d'être joint au set initial et de rendre nos données plus lisibles et compréhensibles.
1. **Première recette** . Suppression des colonnes inutiles et création d'un **indicateur de précision**. Ce dernier est basé sur différents indicateurs à notre disposition et la documentation du dataset.

La documentation contient les informations suivantes :

nbobs_com	Nombre d'observations dans la commune	Un nombre d'observations inférieur à 30 indique une fiabilité faible de l'indicateur de loyer.
nbobs_mail	Nombre d'observations dans la maille	
R2_adj	Coefficient de détermination ajusté du modèle hédonique servant à l'estimation de l'indicateur de loyer	Le coefficient de détermination est d'autant plus élevé que la valeur de l'indicateur est proche des loyers observés dans les annonces. Le R <sup>2</sup> , compris entre 0 et 1, est jugé bon quand sa valeur est supérieure à 0,5

Extrait du tableau de définition des variables issu du **guide d'utilisation du dataset**

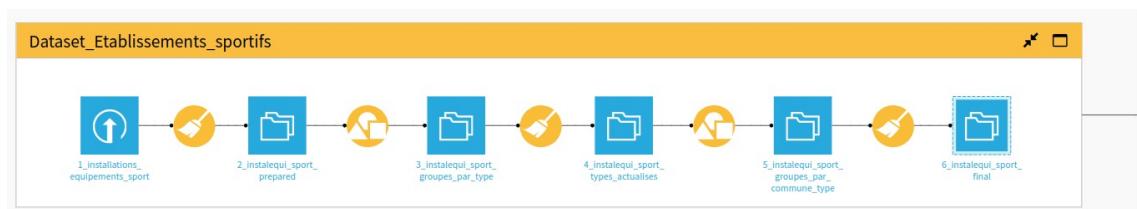
Nous avons utilisé le nombre d'observations dans la commune et le coefficient de détermination pour recenser les communes où la précision de l'indicateur fourni pour les loyer était faible, d'abord pour les maisons puis pour les appartements. Cette évaluation de la qualité de nos données sera probablement réutilisée dans le code de notre application pour indiquer à l'utilisateur un éventuel manque de précision pour une commune donnée.

2. **Deuxième recette** . Nous avons ici renommé les colonnes qui le nécessitaient et supprimé d'autres colonnes inutiles dans le cadre de notre traitement. C'est également ici que nous avons réglé quelques problèmes concernant le format des coordonnées géographiques. Certaines d'entre elles étaient sous la forme « \d.\d E-4 », que nous avons transformé en « 0.0000\d\d ». Nous avons enfin séparé la longitude et la latitude.
3. **Troisième recette** . Cette recette intitulée « Top N » a permis, lorsqu'il y avait des doublons (voir 2.1.1. §« Requêtes SPARQL »), de sélectionner la première ligne d'entre eux par commune après un classement par ordre chronologique (colonne « modificationDate ») pour ensuite supprimer le reste des doublons pour cette même commune.
4. **Quatrième recette** . Jointure entre les résultats de la requête, le dataset des maisons, celui des appartements et celui contenant les libellés des régions et départements.
5. **Cinquième recette** . Jointure entre le dataset des loyers et les enrichissements à partir du code Insee.

6. **Sixième recette** . Nous avons rempli certaines cellules vides et concaténé les noms des départements avec leurs numéro. Nous avons également ajouté le nom du département et de la région pour les communes corses dans la mesure où la Corse était absente du dataset « departements\_regions ». Enfin, nous avons supprimé les colonnes inutiles à la base de données pour avoir un csv ne nécessitant pas de travail supplémentaire à importer dans notre base de données.
7. **Septième recette** . De même que précédemment : remplissage de certaines cellules vides, concaténation des noms des départements avec leurs numéro, ajout du nom du département et de la région pour les communes corses. Dans cette recette, nous avons supprimé les colonnes inutiles au csv servant aux datavisualisations. **C'est ce dernier fichier que nous avons importé dans Tableau.**

## 5.3. TRAITEMENT SUR LES ENRICHISSEMENTS

### 5.3.1. Sport : les équipements sportifs

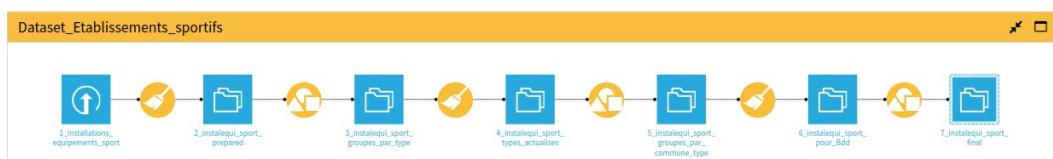


Section du flux consacrée aux équipements sportifs

Le dataset sur les équipements sportifs demande plusieurs étapes de traitements pour aboutir à une table qui soit claire et utilisable dans notre application et une colonne cohérente pour nos datavisualisations.

Le dataset d'origine comprend un grand nombre de données inutiles. Notre objectif premier a donc été de le **réduire** pour ne garder que ce qu'il nous semble utile. Nous sommes partis d'un dataset à **104 colonnes pour aboutir à une table de 3 colonnes** (Insee, type et nombre d'équipements). Cette réduction à l'essentiel s'est faite étape par étape.

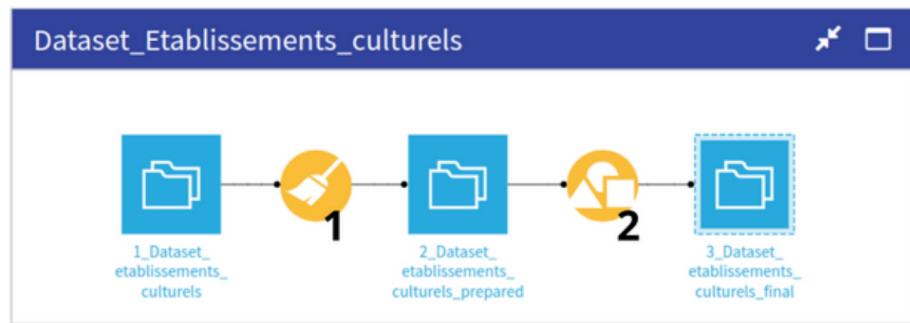
1. **Première recette : suppression des colonnes et renommage.** Sélection des colonnes les plus pertinentes : le code Insee, le numéro et le nom de l'installation ; ainsi que les catégories plus large qui donnent des indications sur la nature de l'équipement (type, famille, Atlas). Nous avons renommé les 6 colonnes restantes pour enlever les majuscules et les accents pour éviter les potentiels bugs.
2. **Deuxième recette : groupement par type d'équipement pour chaque commune en prévision de la recette 3.** Cette étape permet de réunir quelques équipements par leur type pour chaque commune. Une colonne « Count » est créée pour compter les types présents plusieurs fois par commune.
3. **Troisième recette : modification et regroupement des types d'équipements sportifs.** Cette troisième étape a été la plus longue car nous voulions réduire le nombre de type d'équipements de manière cohérente. Nous avions au départ environ 175 types d'équipements différents. Pour réduire ce nombre nous nous sommes basés sur les colonnes de catégories plus larges pour réunir des types sous des appellations plus englobantes. Pour ce faire nous avons utilisé des « Create if then else statement », par exemple si dans la catégorie Atlas il y a la donnée « ski » alors cela renomme les types « ski » (ce qui rassemble les pistes de ski, de luges, les remontées mécaniques etc...). Nous avons supprimé la colonne Famille pour nous baser sur la catégorie la plus large : Atlas.  
Nous avons regroupé sous un même nom des équipements qui étaient enregistrés sous des types différents, comme « terrain de boules » et « terrain de pétanque ». Nous avons aussi supprimé des types qui n'avaient selon nous pas leur place dans le dataset : les Zenith ; les sites de modélisme (la fabrication et le pilotage de modèles réduits).  
Pour finir nous avons renommé les types restants pour obtenir des noms plus explicites que ceux récupérés de la colonne Atlas.  
Tous ces traitements ont permis d'obtenir une bonne réduction de notre nombre de types d'équipements, nous sommes arrivés à 62 types, soit 113 de moins qu'au départ.
4. **Quatrième recette : nouveau groupement par type d'équipements par commune.** Dans cette étape nous répétons l'étape 2 en faisant cette fois un groupement à partir de nos nouveaux types. Cela nous permet d'obtenir dans la colonne « Count » un compte final par commune de chaque type d'équipement sportif.
5. **Cinquième recette : renommage et suppression de colonnes inutiles.** Cette étape nous permet d'obtenir la table finale qui va être utilisée dans notre base de données relationnelle. Nous nettoyons les lignes qui n'ont pas de code Insee, supprimons la première colonne count générée devenue inutile et renommons les 3 colonnes restantes pour qu'elles correspondent aux noms présents dans les autres datasets.
6. **Sixième recette : regroupement final pour la datavisualisation.**



Section du flux consacrée aux équipements sportifs

Cette dernière étape est séparée des autres car elle ne va servir que pour les datavisualisations. Nous faisons un groupement par commune pour avoir le nombre total d'équipements sportifs à intégrer à la table finale. Ce regroupement est effectué sans prendre en compte les différents types d'équipements.

### 5.3.2. Culture : les établissements culturels



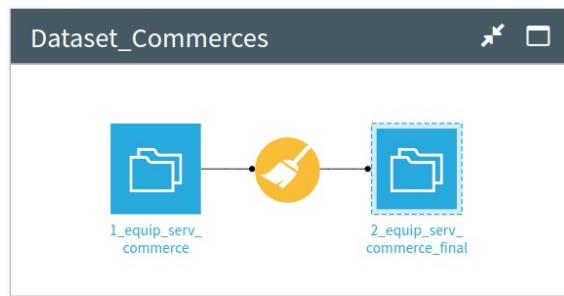
Section du flux consacrée aux établissements culturels

Le jeu de données initial consacré aux données culturelles présente des avantages et inconvénients qui ont pu être traités grâce à dataiku. Nous avons pris le parti de **réduire le volume de ce jeu de données** car de nombreuses catégories sont peu pertinentes pour les utilisateurs de notre application. Par exemple, nous pensons que rares sont les personnes pour lesquelles un monument historique est un argument probant pour emménager dans une commune or ils représentent plus de la moitié de la base : 47 418 monuments sont enregistrés. Au contraire, nous sommes convaincus que les bibliothèques, au nombre de 16 029 au total, sont importantes de par le service au public qu'elles incarnent au quotidien. Puis, nous avons souhaité **offrir à l'utilisateur une vision d'ensemble, et approximative, du nombre d'établissements culturels** par commune. Pour arriver à nos fins, le traitement effectué sur ce jeu de données est le suivant :

- 1. Première recette : la préparation des données.** Ce jeu de données recense un nombre redondant de colonnes alors identiques à celles de notre jeu de données initial. C'est pourquoi nous avons créé une colonne qui réunit le code Insee des communes et ceux des arrondissements des villes de Paris, Marseille, Lyon afin de n'en créer qu'une. Puis, nous avons décidé de garder uniquement les deux colonnes qui nous intéressent soit le code Insee et le type d'équipement du lieu. Nous avons nettoyé les données pour que les codes Insee (qui comportent 5 chiffres) soient complets. Ce à quoi nous avons filtré les lignes pour ne garder que les catégories suivantes : Musée, Cinéma, Théâtre, Conservatoire, Bibliothèque, Scène, Centre de création artistique, Centre d'art, Centre de création musicale, Centre culturel et Opéra. A partir de ce filtre, nous avons pu créer de nouvelles colonnes pour répartir le nombre de ces dernières catégories par code Insee. Pour finaliser cette première recette, nous avons remplacé l'ensemble des données vides par la valeur numérique « 0 » et renommé l'ensemble des colonnes pour clarifier et uniformiser le traitement.
- 2. Seconde recette : grouper les données.** Afin de n'ajouter qu'une colonne à notre table finale, il convient de grouper les données. Ce qui revient à additionner le nombre d'établissements culturels par commune, peu importe leur catégorie précédemment sélectionnée.

Le traitement se résume donc en une succession d'étapes aisément réalisées car les données d'origines sont particulièrement propres et bien formées. Le résultat est autant utilisé pour les datavisualisations que pour la base de données de l'application.

### 5.3.3. Commerce : un regroupement par catégories



Section du flux consacrée aux commerces

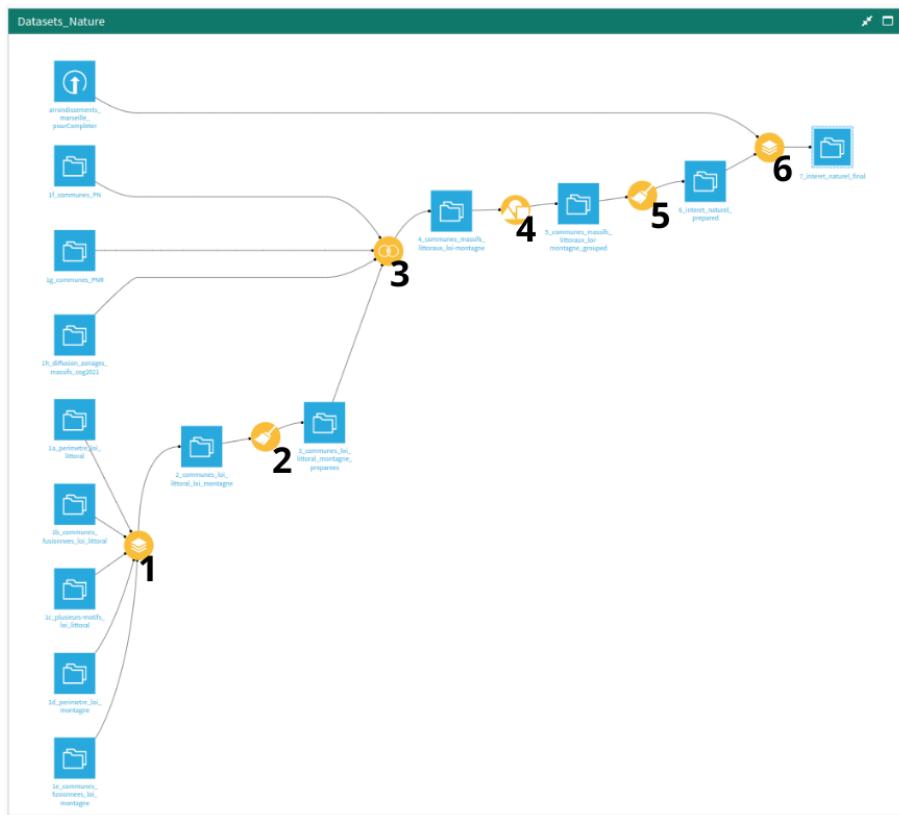
Après avoir examiné ce jeu de données sur le commerce, nous avons décidé d'explorer une dimension particulière du nettoyage de données : l'introduction d'une nouvelle étape d'organisation. Bien que les variables soient pertinentes, elles ne répondent pas pleinement à nos exigences. En l'absence de regroupement, ces variables présentent une multitude d'informations spécifiques, moins communicatives pour les utilisateurs susceptibles de s'intéresser à des données de visualisation plus intuitives.

**Recette unique.** La première et unique étape consiste à regrouper les différentes variables dans des catégories spécifiques et à appliquer une somme à chacun des catégories.

- la première catégorie « Alimentation » : Hypermarchés, Supermarchés, Supérettes, Épiceries, Boulangeries, Boucheries charcuteries, Produits surgelés et Poissonneries ;
- la deuxième catégorie « Commerce général » : Magasins de vêtements, Magasins d'équipements du foyer, Magasins de chaussures, Magasins d'électroménager et de matériel audio-vidéo, Magasins de meubles, Magasins de revêtements murs et sols, Grandes surfaces de bricolage, Magasins d'optique, Magasins de matériel médical et orthopédique, ainsi que les Droggeries quincailleries bricolages ;
- la troisième catégorie « Loisirs » : Librairies, Magasins d'articles de sports et de loisirs ;
- la quatrième catégorie « Beauté et accessoires » : Parfumeries - Cosmétiques et Horlogerie-Bijouteries ;
- la cinquième catégorie « Jardinage et Animalerie » : Fleuristes, Jardineries et Animaleries ;
- la sixième catégorie « Services » : Stations-services.

Finalement, après avoir organisé ces variables dans des catégories spécifiques, nous avons effectué une deuxième somme, représentant le total de toutes les variables, pour obtenir le nombre total d'unités commerciales sur le territoire français.

### 5.3.4. Nature : les communes d'environnement protégé



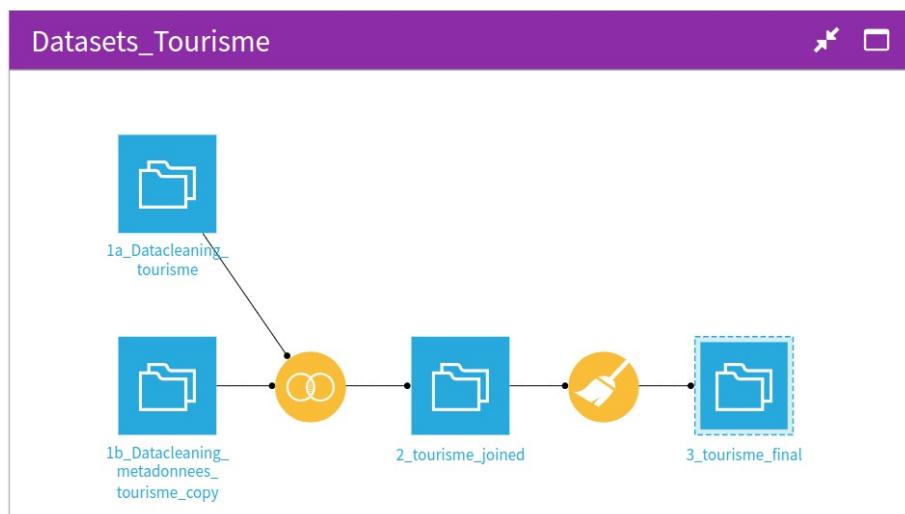
Section du flux consacrée à l'environnement naturel

0. **Datasets** . Nous avons importé les différents datasets pour en faire une seule table à la fin du flux. Les datasets de La loi Littoral et de la Loi Montagne en tout sont visibles dans le flux pour ces datasets.
1. **Première recette** . Regroupement de l'ensemble des datasets sur les lois Montagne et Littoral pour ensuite pouvoir les traiter. Il s'agit de fichiers Excel divisés en plusieurs feuilles. Le dataset Loi Littoral présente une feuille générale contenant les communes de France, une feuille contenant les communes fusionnées et une feuille consacrée aux communes concernées par plusieurs motifs (qui peuvent être soumis à la loi pour un estuaire et la mer par exemple). Le fichier excel Loi Montagne contient la feuille générale et celle sur les communes fusionnées. Nous avons donc dû importer et empiler cinq fichiers dans notre flux.
2. **Deuxième recette** . Préparation des datasets Loi Montagne et Loi Littoral pour la jointure avec celui sur les massifs : nous avons créé des colonnes contenant des booléens (dans un premier temps sous la forme 0/1) pour les communes dont le classement est issu de la proximité avec la montagne, la mer, un estuaire ou un lac.

3. **Troisième recette** . Jointure avec les fichiers contenant les données sur les PN, sur les PNR, et sur les zones de massifs grâce au code Insee.
4. **Quatrième recette** . Groupement par code Insee afin de supprimer des doublons. Ces derniers ont soit été générés parce que les communes sont issues d'une fusion, soit parce qu'elles sont concernées par plusieurs arrêtés de la Loi Montagne.
5. **Cinquième recette** . Préparation des données afin de les rendre plus lisibles. Les booléens ont notamment été transformés pour plus de clarté en « True » et « False » et les cellules vides dans ces colonnes booléennes ont été remplies avec la valeur « False ».
6. **Sixième recette** . Notre dataset sur les loyers concerne les communes et arrondissements, or les datasets traités pour l'environnement géographique concernent les communes. Nous avons choisi de créer un csv contenant les informations sur les arrondissements de Paris, Lyon et Marseille en se basant sur la ligne respective de chaque commune, pour le joindre au dataset final sur la nature.

L'indicateur sur la nature répond aux besoins de notre application et n'a pas été utilisé pour produire les datavisualisations.

#### 5.3.5. Tourisme : les hébergements touristiques



Section du flux consacrée aux hébergements touristiques

Ce jeu de données a suscité notre attention en raison de problèmes liés aux faux positifs, notamment concernant les campings et les emplacements de camping. Bien que ces éléments soient essentiels pour l'application, ils semblent moins pertinents pour la datavisualisation, car ils sont souvent sujets à des fluctuations. Ces valeurs ne reflètent pas nécessairement la réalité des hébergements. De plus, nous avons entrepris un processus de nettoyage, en identifiant les variables pertinentes pour rendre compte de la capacité d'hébergement en lien avec le tourisme. Étant donné que la quantité d'hôtels ne traduit pas toujours la capacité d'un hébergement, nous avons également choisi la quantité de chambres comme deuxième variable afin de gérer d'éventuelles incohérences.

- 1. Première recette .** Nous avons réalisé une jointure (left join) entre le tableau contenant les données sur le tourisme (1a\_Datacleaning\_tourisme) et le tableau contenant les noms des communes contenues dans ce dernier (1b\_Data-cleaning\_metadonnees\_tourisme\_copy). Ils sont tous deux issus du dataset de départ. Cette liaison s'est opérée à travers la clé « COD\_MOD » du deuxième ensemble de données, qui correspondait au « code\_geo » dans le premier ensemble de données.
- 2. Deuxième recette .** La deuxième étape de notre procédure a consisté en le traitement des colonnes, impliquant une modification de leur ordre, des renommages, et la somme des valeurs de certaines colonnes. Les colonnes sélectionnées pour générer cette somme d'hébergements touristiques, présente dans le dernier ensemble de données « 3\_tourisme\_final », sont : la somme des hôtels, des chambres d'hôtels, des villages vacances/maisons familiales, des résidences de tourisme et des auberges de jeunesse/centres sportifs pour l'année 2023, à l'exception des campings et des emplacements de camping.

### 5.3.6. Conclusion

Pour conclure, il nous a fallu appréhender l'outil dataiku pour réaliser le travail. Certains membres du groupe habitués à utiliser le tableur Excel ont finalement compris toute la force de cette plate-forme de traitement des données. Dans l'ensemble, l'utilisation de Python ou encore des expressions régulières ont été d'une grande utilité. Le traitement des données a étonnamment été d'une grande facilité pour certains dataset (culture, commerces) et inversement d'une plus grande difficulté pour d'autres (sport, nature). L'exportation et l'importation du flux d'un ordinateur à un autre n'a pas posé de problème principal.

# DATAVISUALISATIONS

## 6. PRÉAMBULE

### 6.1. AVANT-PROPOS

La dernière partie du traitement de données a dupliqué le dataset final fournissant un fichier pour la base de données et un autre pour les datavisualisations, qui sont l'objet de cette nouvelle et dernière partie du journal de bord.

#### 6.1.1. Nature des données pour la datavisualisation

Avant de réaliser les visualisations de données, il convient de définir la nature de ces dernières. Ci-dessous, une liste qui les présente :

- INSEE\_C : donnée qualitative nominale
- LIBGEO : donnée qualitative nominale
- LOYERM2\_MAISON : donnée quantitative continue
- LOYERM2\_Appart : donnée quantitative continue
- SUPERFICIE : donnée quantitative continue
- POP : donnée quantitative continue
- LAT : donnée quantitative repérable
- LONG : donnée quantitative repérable
- DEP : donnée qualitative nominale
- REGION : donnée qualitative nominale
- INTERET\_NATUREL : donnée qualitative nominale
- NBRE\_CULTURE : donnée qualitative continue
- NBRE\_SPORT : donnée qualitative continue
- HERBEGEMENTS\_TOUR : donnée qualitative continue
- NBRE\_COMMERCES : donnée qualitative continue

Cette dénomination a facilité la visualisation des données pour comprendre plus facilement quelles sont les données comparables entre elles.

# 7. LES DATAVISUALISATIONS

## 7.1. ILLUSTRATION ET ANALYSE

Chaque datavisualisation (dynamique ou statique) listée ci-dessous est accompagnée d'une analyse. Nous invitons le lecteur à cliquer sur les liens correspondants aux datavisualisations pour lire en parallèle les deux productions : l'analyse textuelle et la présentation visuelle des données.

### 7.1.1. Nuages de points : Indicateur du prix des loyers pour les maisons et les appartements par commune, par département, en fonction ou non de la densité de population

#### Voir et interagir avec la datavisualisation



Nuages de points : Indicateurs des prix des loyers ( $\text{€}/\text{m}^2$ ) en France par commune, en fonction ou non de la densité de population.

- Les données utilisées : les **prix des loyers au  $\text{m}^2$  pour les maisons** et les appartements à partir du dataset initial, la **population** et la **superficie** des villes.
- Les champs calculés : Calcul de la densité de population pour chaque commune à partir de la superficie et population

### Une vision globale des données sur les loyers

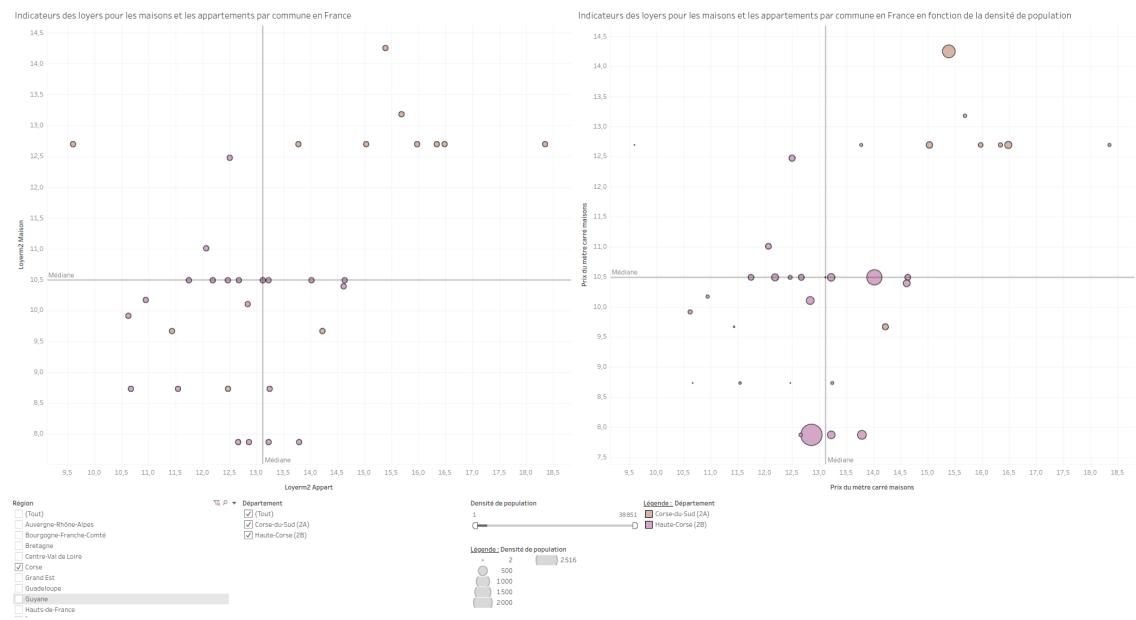
Le nuage de points est un type de datavisualisation relativement simple à effectuer qui est par conséquent souvent utilisé pour avoir une **vision globale** de la tendance de certaines données. On y perçoit en effet facilement le **minimum** et le **maximum** des indicateurs. Dans notre cas, le minimum de l'indicateur du prix du mètre carré pour les loyers des maisons et des appartements se situe autour de 5€ et le maximum autour de 30€. Les points sont davantage regroupés vers le bas de la diagonale que vers le haut. La majorité des villes présente un indicateur entre 5 et 15€ par mètre carré. Les loyers très chers font donc figure d'exception. Ils concernent principalement l'Île-de-France. Ces informations ne sont visibles qu'en quelques coups d'œil. C'est ce qui fait l'intérêt de ce type de diagrammes.

Le nuage de points est également un bon moyen pour repérer rapidement les **faiblesses et éventuelles aberrations dans les données**. Sa réalisation nous a en effet permis de repérer les doublons dans la requête SPARQL abordée précédemment, que nous avons ensuite supprimés. Une autre faille dans les données est bien visible sur le diagramme. Il s'agit de la question des « **mailles** ». En effet, on peut remarquer des lignes verticales ou horizontales de points sur le graphique. Il s'agit de villes qui ont le même prix au mètre carré pour les maisons mais un prix différent pour les appartements et inversement. Ces communes font en réalité partie de la même maille de données. Dans notre dataset de départ, une maille est un regroupement de communes réalisé grâce à un algorithme de clustering spatial. Il s'agit d'une association de données concernant des communes dont les informations n'étaient pas suffisantes pour déterminer le prix du mètre carré. Les mailles sont très visibles concernant Paris parce que les prix des loyers pour les appartements sont élevés et se démarquent ainsi vers le haut du nuage. Si l'on s'intéresse à la densité de population, ajoutée au graphique de droite comme attribut déterminant la taille des points à notre graphique, on se rend compte que les lignes horizontales de points concernent en général des villes densément peuplées. Il s'agit également de villes touristiques littorales.

On peut émettre l'hypothèse selon laquelle les prix des loyers pour les appartements et les maisons y sont différents car l'offre pour les appartements est supérieure à celle pour les maisons. Les fortes densités et le coût élevé des maisons à ces endroits pourraient contribuer au fait qu'il y ait peu d'offre pour les maisons, qui sont davantage mises en vente que louées. Inversement, les lignes verticales correspondent plutôt à des zones où il y a peu d'offres pour les maisons et donc à des zones où les densités sont plus faibles. Les villes où il y a peu d'offre de location en général sont quant à elles rassemblées en un point, et leur quantité est donc invisibilisée sur le graphique. Ce nuage de points nous rappelle ainsi les failles de notre dataset initial, qui peut parfois manquer de précision lorsqu'il y a peu d'offres de location.

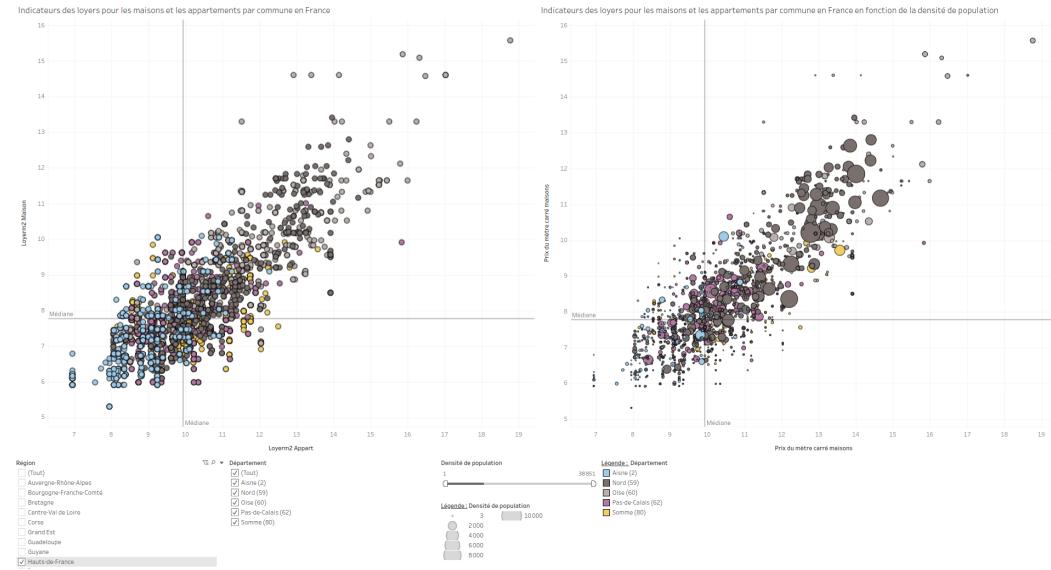
L'exemple de la Corse, illustré ci-dessous, est assez parlant : peu de points sont représentés, ce qui sous-entend que de nombreux points se superposent, et, quand ces derniers ne se superposent pas, ils sont représentés sous la forme d'une ligne droite. C'est un territoire dont les données sont majoritairement sous forme de mailles car il y a peu d'offre de location mensuelle et certainement plus d'offres d'achat ou de locations touristiques de courts séjours.

# DATAVISUALISATIONS



Nuages de points filtré : Indicateurs des prix des loyers ( $\text{€}/\text{m}^2$ ) en Corse par commune, en fonction d'un filtre sur la densité de population

## Une datavisualisation dynamique pour établir des comparaisons



Nuages de points filtré : Indicateurs des prix des loyers ( $\text{€}/\text{m}^2$ ) dans les Hauts-de-France par commune, en fonction ou non de la densité de population

Des **filtres** offrent la possibilité de n'afficher que certaines densités de populations ou de ne sélectionner que des départements ou régions et pour ainsi obtenir une vision à une échelle plus locale. La réduction de l'échelle des densités permet par exemple de masquer les villes très densément peuplées pour mieux voir les points qui sont très petits quand aucun filtre n'est utilisé. L'utilisateur peut aussi choisir d'**effectuer des comparaisons** au sein d'une région ou entre deux départements. Le choix d'une seule région permet d'avoir une meilleure visibilité sur les similarités en termes de prix des loyers au sein des départements, qui forment des masses assez uniformes. Si l'on prend l'exemple de la région Hauts-de-France, les prix sont plus élevés dans le Nord et l'Oise. On peut émettre l'hypothèse que la métropole lilloise, densément peuplée et attractive, dans le Nord et le fait que l'Oise soit proche de Paris ont une influence sur ces prix plus élevés mais il nous faudrait faire davantage de recherches pour la confirmer. Les comparaisons permettent de souligner la **diversité des loyers selon les départements**. Sur le diagramme, les prix maximaux sont en effet jusqu'à six fois supérieurs aux prix minimaux en France. Il existe également une certaine **diversité au sein des départements**. Les points concernant le département du Nord paraissent très étalés. Les points les moins hauts sont entre 6 et 7 $\text{€}$  pour un mètre carré pour les appartements ou les maisons, alors que le plus haut se situe entre 13 et 14 $\text{€}$ , soit le double.

Ce graphique rend ainsi compte de la diversité des prix des loyers sur le territoire français, qui paraissent plus chers dans les zones touristiques, de littoral et de fortes densités, et inversement moins chers dans les zones peu densément peuplées.

### Quelques failles

Comme mentionné précédemment, nos données ne sont pas parfaites. Leur méthode de récupération favorise les biais. Il faut donc être prudent en interprétant ce graphique. Cette datavisualisation présente également des inconvénients dans la mesure où elle est **très générale**. Le nombre élevé des couleurs peut être déroutant si tous les filtres sont sélectionnés. Le nombre de départements est trop élevé pour la palette de couleur utilisée. Certaines couleurs concernent donc deux départements. De plus, les points se superposent et ne rendent pas compte de la concentration réelle de points vers le bas du diagramme et de l'étendue des villes qui ont le même prix au mètre carré pour les maisons et les appartements.

La **médiane** sépare les deux moitiés des séries de valeurs des deux axes, son affichage montre à quel point il est difficile de se rendre compte des densités de points, qui sont ici très grandes vers le bas du graphique, puisqu'il y a trop de données à afficher. Une visualisation en 3D pourrait être intéressante pour mieux rendre compte de la superposition des points. Dans le cas de notre nuage de points, l'utilisateur ne doit se concentrer que sur quelques départements. Les couleurs sont également faiblement lisibles parce que certaines sont utilisées pour plusieurs départements. En termes de superposition, les points sont ordonnés par ordre alphabétique, les points des premiers départements dans l'alphabet sont donc davantage visibles.

En ce qui concerne les interprétations, n'étant pas des spécialistes en économie ou en démographie, nous pouvons émettre des hypothèses à partir de ce diagramme, par exemple, la densité de population semble avoir une influence sur le prix des loyers dans les zones les plus hautes ou les plus basses du graphique. Toutefois, nous ne sommes pas en mesure de les valider. Nous pouvons simplement affirmer que le nuage de points souligne une certaine **diversité en termes de loyers au sein du territoire français**, ce qui rappelle que la France est un pays d'une grande complexité. Cela fait également écho au fait que **son découpage territorial n'est pas forcément uniforme**. Les communes sont de superficies, populations et budgets différents, de même pour les départements et les régions.

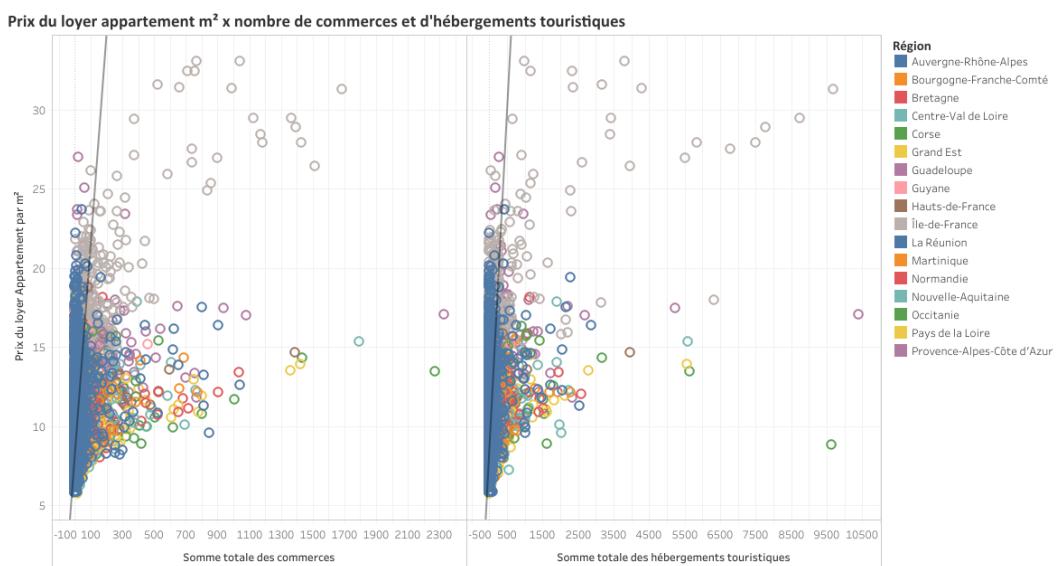
### 7.1.2. Relation entre les loyers des appartements et maisons en fonction de la somme totale des commerces et hébergements touristiques

[Voir la datavisualisation pour les appartements](#)

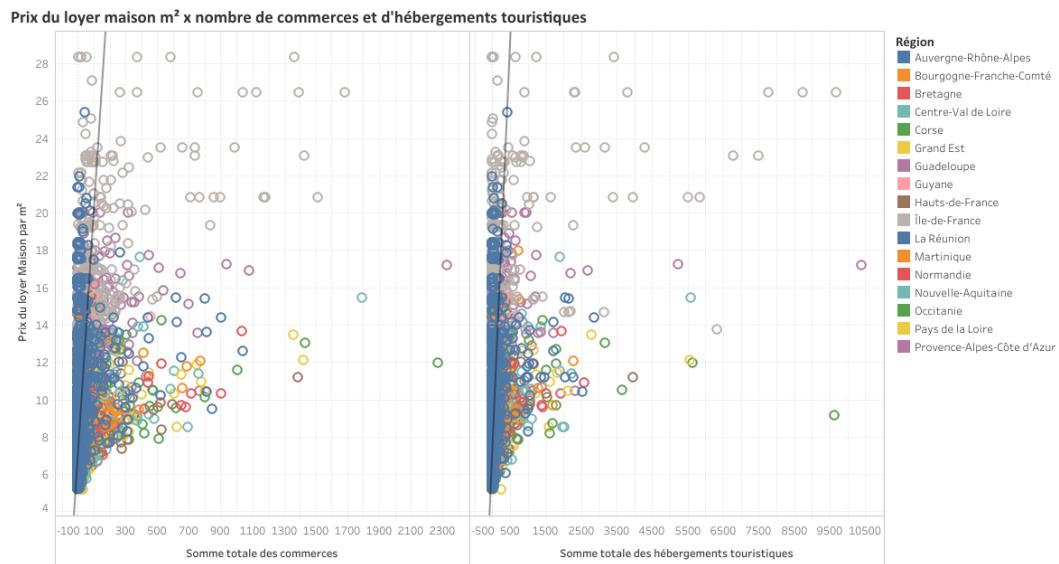
[Voir la datavisualisation pour les maisons](#)

Un second graphe de points a été généré dans le but **d'explorer les relations entre les loyers d'appartements et de maisons** d'une part, et les secteurs du commerce et du tourisme d'autre part. Dans cette démarche, les variables indépendantes sont définies comme la somme totale de commerces et la somme totale d'hébergements touristiques. Notre objectif était d'analyser l'influence de ces variables sur les prix des appartements et des maisons.

Pour accomplir ce projet, nous avons commencé par **établir des relations entre les prix d'appartements et, par la suite, sur les prix des maisons**. En prenant une définition de variable dépendante et indépendante, nous pouvons dire que la variable dépendante est la caractéristique mesurée ou observée dans une étude, tandis que la variable indépendante est la variable manipulée ou influençant la variable dépendante. Les variables indépendantes dans notre étude sont la somme totale de commerces et la somme totale d'hébergements touristiques. Nous avons choisi ces variables dans le but de comprendre comment elles modifient ou impactent les prix du loyer. **La quantité de commerces est positionnée sur l'axe des x (horizontal) tandis que les prix du loyer sont représentés sur l'axe des y (vertical) dans nos analyses graphiques**. Ces choix permettent d'explorer visuellement la relation entre ces variables et d'évaluer comment les variations dans la quantité de commerces sont associées aux fluctuations des prix du loyer.



Indicateurs des prix des loyers des appartements (€/m<sup>2</sup>) en fonction de la somme totale des commerces et des hébergements touristiques

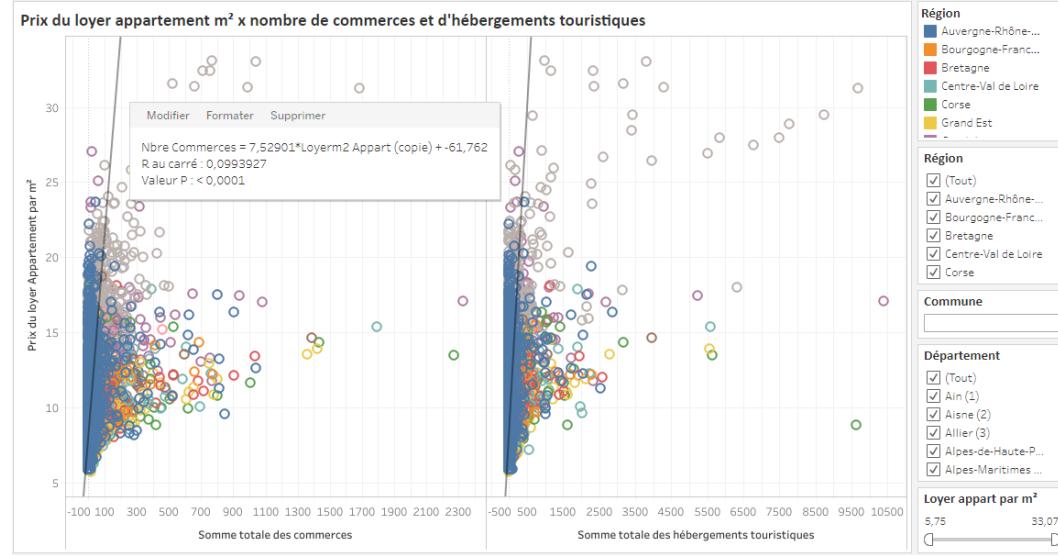


Indicateurs des prix des loyers des maisons (€/m<sup>2</sup>) en fonction de la somme totale des commerces et des hébergements touristiques

Il est essentiel de souligner que notre démarche ne vise pas une analyse rigoureuse à la manière d'une étude scientifique. Cette approche requerrait une maîtrise approfondie des concepts statistiques ainsi qu'une capacité à bien comprendre les éventuels biais associés à *TableauPublic*. Notre objectif principal consiste plutôt à observer et à présenter les informations qui nous sont fournies.

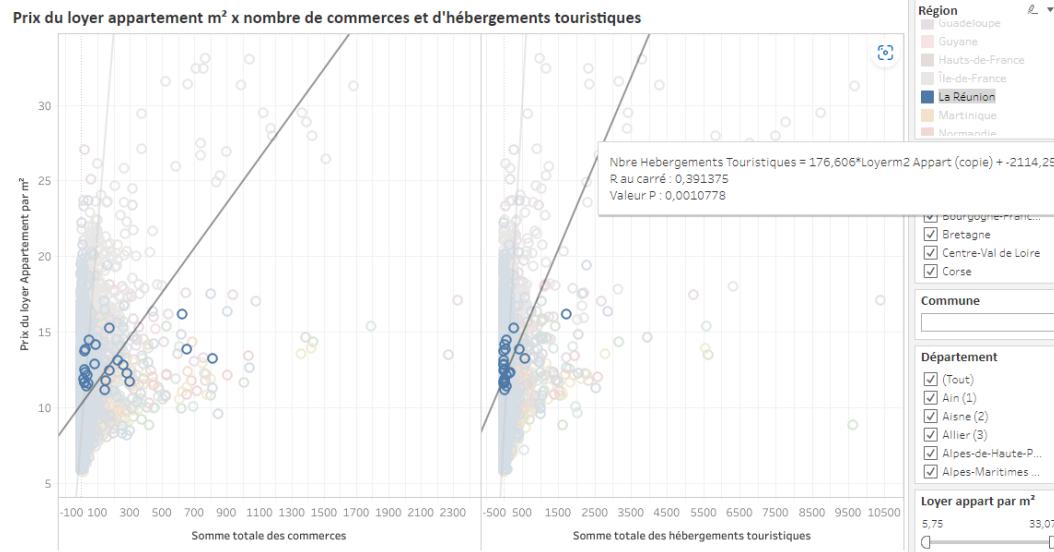
Afin d'accomplir cette tâche, nous avons employé la **régression linéaire**, l'un des outils disponibles sur *TableauPublic*. La régression linéaire est utilisée pour **modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes**. L'objectif est de créer une **équation linéaire** qui représente au mieux cette relation, permettant ainsi de comprendre et de prédire le comportement de la variable dépendante en fonction des variables indépendantes. Cette approche est largement appliquée dans divers domaines tels que l'économie, la biologie et la finance pour analyser les tendances, évaluer les effets et formuler des prédictions basées sur des données observées. Cette équation linéaire prend la forme  $y = ax + b$ , où  $y$  est la variable dépendante,  $x$  est la variable indépendante,  $a$  est la pente de la ligne, et  $b$  est l'ordonnée à l'origine. Le coefficient de détermination, souvent noté  $R^2$ , mesure la proportion de la variance de la variable dépendante qui peut être expliquée par le modèle de régression. Un  $R^2$  proche de 1 indique que le modèle explique une grande partie de la variation des données, soulignant ainsi sa pertinence. Cependant, il est important de noter que **l'interprétation d'un  $R^2$  doit être complétée par d'autres mesures** et une analyse approfondie pour assurer une évaluation complète du modèle de régression linéaire. Il est encore possible qu'une régression puisse avoir un  $R^2$  très haut, mais cela ne veut pas dire non plus qu'elles expriment une causalité.

Quoi qu'il en soit, il est essentiel de souligner que *TableauPublic* met à notre disposition la courbe de tendance en tant qu'outil d'analyse dans notre étude. En examinant les résultats fournis par Tableau pour notre ensemble de données, nous observons, par exemple, un coefficient de détermination  $R^2$  avec une valeur approximative de 0,09. En matière de régression linéaire, cela pourrait **indiquer une corrélation presque inexisteante entre les variables indépendantes et dépendantes**, étant donné que  $R^2$  est considérablement éloigné de 1. Cela **suggère également que l'analyse de nos données ne peut pas se limiter à ce seul critère**.



Régression linéaire :  $R^2$  des prix des loyers des appartements ( $\text{€}/\text{m}^2$ ) en fonction de la somme totale des commerces

En examinant le nuage de points dans le contexte de la région française de La Réunion, nous constatons toutefois un  $R^2$  d'environ 0,39, avec davantage de points se rapprochant de la ligne. Une valeur de  $R^2$  plus proche de 1 suggère que la ligne de régression (la courbe de tendance) capture efficacement la variation des données. Autrement dit, la plupart des points dans le nuage sont relativement proches de cette ligne de régression.



R au carré des prix des loyers des appartements dans La Réunion en fonction de la somme totale des hébergements touristiques

### 7.1.3. Diagramme en barre : Comparaisons loyers/equipements par région

[Voir la datavisualisation](#)

Un diagramme en barre permet de faire des comparaisons entre différentes catégories. Ce type de graphique est très lisible et est donc un outil visuel intéressant lorsqu'on souhaite aider l'utilisateur dans la compréhension des relations et des différences entre plusieurs catégories de données.

Pour cette visualisation l'idée était de faire une comparaison, à l'échelle régionale, entre l'indicateur de loyer et les divers équipements dont disposent les régions. **Est-il possible de voir une corrélation entre un loyer élevé et une région très bien équipée ?**

Les équipements choisis sont les équipements sportifs, les établissements culturels et les commerces, des éléments qui pourraient éventuellement impacter le montant des loyer. Pour chaque catégorie nous avons effectué un calcul pour rapporter la quantité d'équipement à la population de chaque région : nous nous sommes basés sur le nombre de personnes pour un équipement. (NB : ce calcul peut prêter à confusion dans la lecture du diagramme, **plus la barre est petite, mieux la région est équipée** par rapport à sa population. En effet le calcul inverse proposait des données trop petites ce qui rendait le diagramme illisible pour certaines régions)

Pour les loyers, une moyenne a été faite entre les prix pour les maisons et pour les appartements pour obtenir une donnée générale. Nous avons ajouté un total général pour toutes les catégories pour ensuite produire une ligne de moyenne à laquelle se référer pour chaque région.

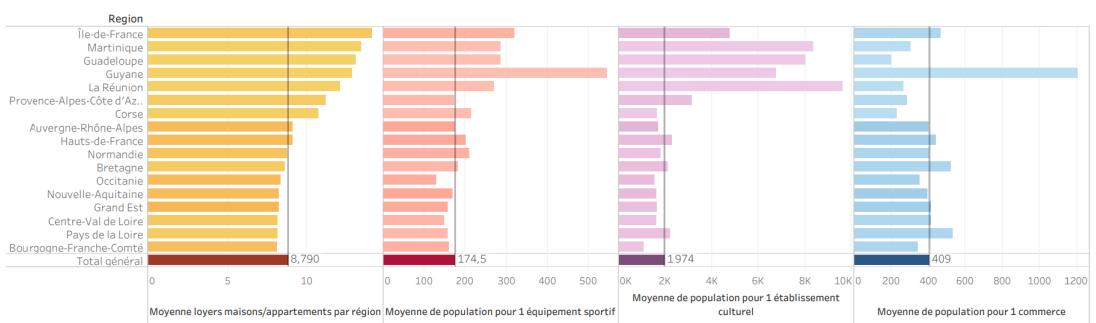
**En ordonnant les loyers du plus élevé au plus bas il est clair que les autres catégories ne suivent pas.**

Les régions aux loyers plus élevés semblent avoir moins d'équipements sportifs et culturels par personne par exemple. Pour l'Île-de-France cela pourrait s'expliquer par une population très élevée qui ne compenserait pas l'offre culturelle pourtant très élevée elle aussi. Pour les régions d'outre-mer le loyer est très élevé mais il y peu d'équipements par personne. La Guyane en particulier nous a surpris car il semble qu'il y a très peu de commerces par personne.

Quant aux régions qui ont les loyers les plus bas comme la Bourgogne-Franche-Comté ou les Pays de la Loire, il semble qu'elles soient mieux équipées que l'Île-de-France dans la plupart des catégories.

Cette première lecture est intéressante mais il faut garder en tête quelques nuances, **les régions ne font pas toutes la même taille et n'ont pas la même densité de population.** Analyser les données à cette échelle est donc à faire avec précaution.

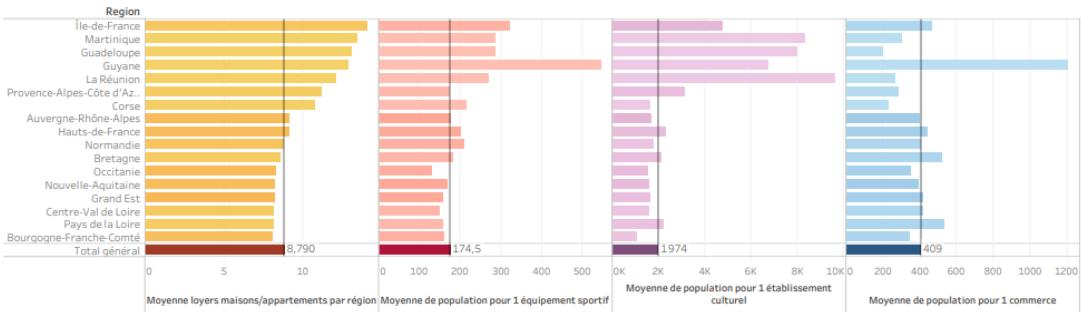
Comparaison par région Moyenne



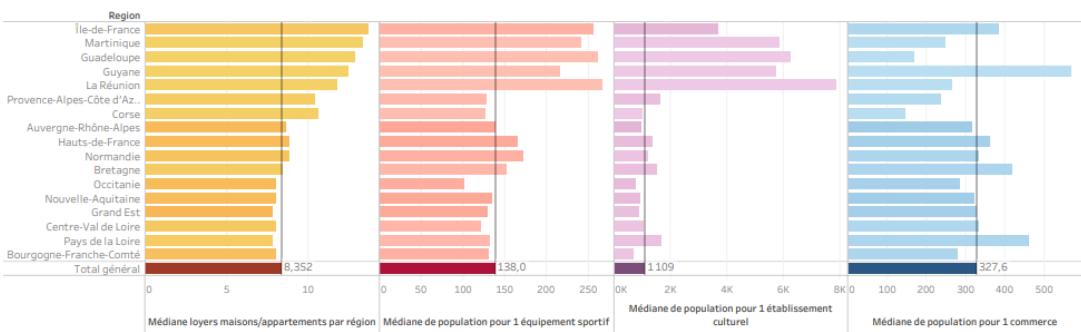
Comparaison loyers-equipements par région. Moyenne

**Citynder - Comparaisons par régions.**  
 Ces deux diagrammes en barre comparent les loyers par région à la quantité d'équipements dont la région dispose. Dans le premier les données sont calculées en moyennes tandis que dans le second elles sont calculées en médianes. Une ligne de constante a été créée pour permettre de situer chaque région au dessus ou en dessous de la moyenne générale dans toutes les catégories présentées.

Comparaison par région Moyenne



Comparaison par région Médiane



### Comparaison loyers-équipements par région. Moyennes et Médianes

Un autre élément à prendre en compte est la différence qu'il peut y avoir entre les résultats en fonction du type de données utilisées. C'est pour cette raison que dans le tableau de bord nous avons gardé la version calculée avec des médianes et celle avec des moyennes. Cette comparaison permet de voir les différences de résultats. C'est surtout un changement d'échelle qui est visible avec la version des médianes qui a des données plus basses. La médiane divise les données en deux parts égales, alors que la moyenne est la somme des données, divisée par le nombre de celles-ci. La médiane est donc un point central, elle permet d'éliminer les valeurs extrêmes et d'exprimer la valeur du milieu.

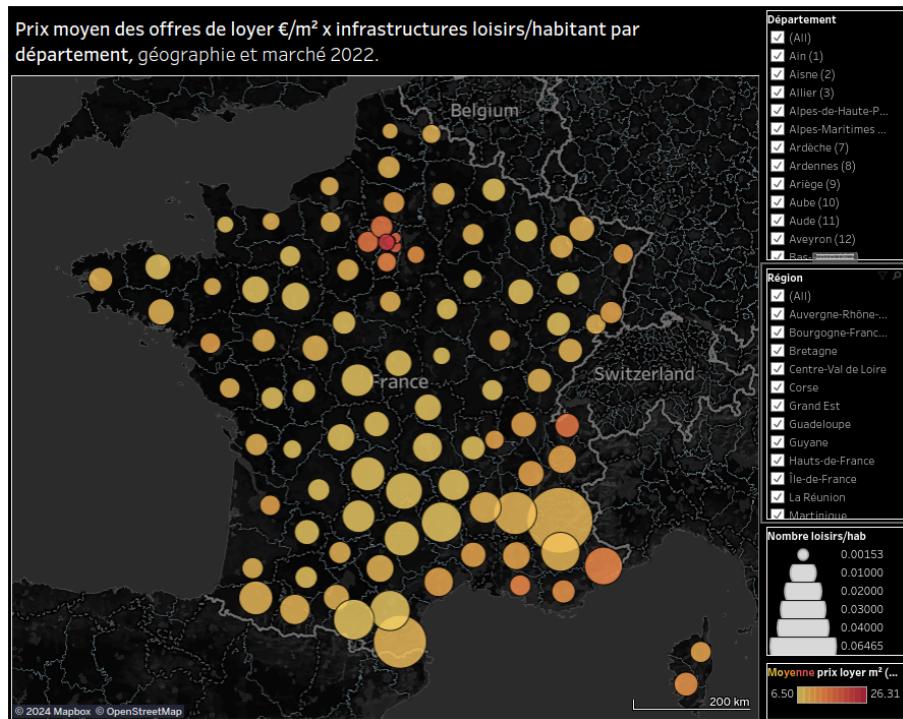
**Cette médiane permet donc de nuancer les résultats** pour les régions d'outre-Mer par exemple, même si les tendances restent similaires aux résultats obtenus avec les moyennes.

Malgré ses limites, ce diagramme nous permet de proposer des hypothèses quant aux liens qui existent ou non, entre les loyers et les équipements. Ici, il semble qu'avec nos données, il soit impossible d'affirmer une corrélation entre les catégories, mais ce n'est qu'une hypothèse proposée en ayant conscience des limites de nos données ainsi que de nos capacités à analyser celles-ci (nous ne sommes ni statisticiens, ni économistes).

Ce diagramme nous permet surtout d'avoir une vision globale de la diversité des situations en France, et de commencer à amorcer des réflexions à partir de ces résultats.

#### 7.1.4. Carte interactive : Prix moyen des offres de loyer €/m<sup>2</sup> et infrastructures Loisirs/habitant par département, selon la géographie et le marché de 2022.

[Voir et intégrer avec la datavisualisation](#)



Carte : prix moyen des offres de loyer (maison et appartement) au m<sup>2</sup> et infrastructures "loisirs" par habitant pour les départements de France hors Mayotte (capture d'écran centrée sur la France métropolitaine dont la Corse).

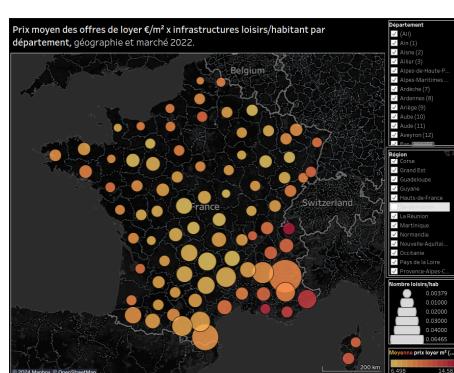
#### Présentation des choix arbitraires

- La carte : utiliser une carte pour visualiser des **données géographiques** nous semble essentiel. Non seulement parce que la commune est une division administrative du **territoire**, et, le point central de notre jeu de données, mais aussi, et surtout, car nous disposons pour chacune d'elle, des coordonnées longitude et latitude.
- L'échelle : régionale, départementale ou communale ? Depuis la création des nouvelles régions métropolitaines (2016), trois régions en France (Nouvelle-Aquitaine, Auvergne-Rhône-Alpes, Occitanie) sont dotées d'une superficie disproportionnée par rapport aux autres régions de la métropole (Île-de-France, Bretagne, Corse). Par conséquent, il paraît douteux, au premier abord, de les comparer entre elles. A l'échelle de la commune, le propos semble le plus pertinent pour représenter la donnée primaire de jeu : la commune. Cependant, les villes par arrondissements (Paris, Marseille, Lyon) nous ont posé quelques problèmes (expliqués plus loin). C'est pourquoi **l'échelle départementale** est celle qui a présenté le plus d'arguments convaincants : réduction du nombre de points, espacement des données spatialement réparties, clarté du propos illustré, etc.

- Les données sélectionnées :
  - le dataset initial : le **prix du loyer au m<sup>2</sup> pour les maisons et les appartements**. Nous avons tous convenu que la datavisualisation doit *a minima* aborder le sujet du loyer car c'est la donnée fondamentale de notre travail.
  - les enrichissements : les **infrastructures sportives et établissements culturels**. Ils ont été privilégiés pour cette datavisualisation car il s'agit de données comparables par leur nature (donnée qualitative continue). Cependant, la proportion de chaque enrichissement est foncièrement différente : alors que le total d'établissements culturels choisis pour notre travail avoisine les 21K, celui des infrastructures sportives est de 328K. Il faut donc avoir conscience que les données sportives sont proportionnellement supérieures aux données culturelles. Cependant, nous pensons qu'ils peuvent être réunis sous la catégorie «**Loisirs**».
- Les champs calculés : réunir les données pour n'en former qu'une.
  - Calcul de la moyenne du prix des loyers = (**[Loyer m<sup>2</sup> Appart] + [Loyer m<sup>2</sup> Maison]**) / 2)
  - Calcul de la somme des infrastructures loisirs = (**[Nbre Equipements Sportifs] + [Nbre Etablissement Cul]**) / 2)
  - Calcul du nombre des infrastructures par habitant = (**[Somme des infrastructures loisirs] / [POP]**)

A partir de ces calculs, **le prix moyen des offres de loyer €/m<sup>2</sup> est représenté par une couleur graduée (du jaune au rouge) et le rapport d'infrastructures "Loisirs" par habitant est représenté par la taille du cercle** alors coloré. Chaque donnée est filtrée par département - son point géographique de référence étant représenté par sa capitale administrative.

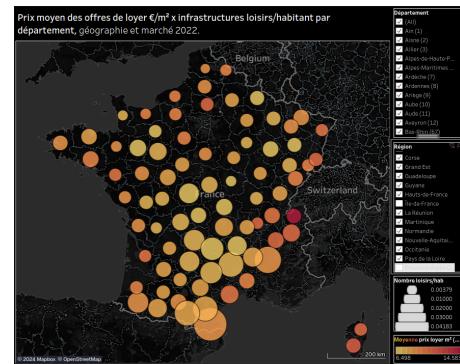
Pour accompagner l'analyse de cette datavisualisation, nous invitons le lecteur à intéragir avec cette dernière pour **naviguer sur le globe** et observer le résultat pour les **départements d'outre-mer**, qui ne peuvent être encapsulés à côté de la Métropole dans la datavisualisation finale. Pour l'ensemble des niveaux de lecture proposés ci-dessous, il est conseillé de **décocher les régions**, dans cet ordre et au-fur-et-à-mesure, **d'abord l'Île-de-France, puis ensemble la Provence-Alpes-Côtes-d'Azur, l'Auvergne-Rhône-Alpes, l'Occitanie et la Corse**.



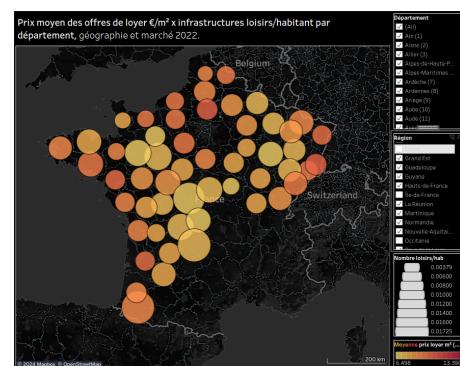
Carte : prix moyen des offres de loyer (maison et appartement) au m<sup>2</sup> et infrastructures "loisirs" par habitant sur l'ensemble du territoire métropolitain sans l'Île-de-France (capture d'écran centrée sur la France métropolitaine dont la Corse).

**Plusieurs niveaux de lecture :**

- Les loyers : de par la couleur des cercles, nous voyons que les loyers les plus élevés sont rapprochés de la **capitale française**. Puis une deuxième zone se révèle être aussi dans la moyenne supérieure des prix : de la **côte méditerranéenne jusque dans les terres vers les Hautes-Alpes**, et la Corse. Le reste de la France semble plus ou moins homogène. Toutefois, nous pensons que les données extrêmement hautes écrasent le reste des données inférieures, extrêmement basses ou plutôt dans la moyenne, rendant le reste assez homogène car bien inférieur à l'extrême. Cependant, lorsque nous retirons le département de Paris (comprenant entièrement les arrondissements de Paris), les données s'actualisent et nous constatons que cette deuxième zone est en effet la seconde plus élevée de France. Elle comprend ainsi les départements des Bouches-du-Rhône (majoritairement représentée par Marseille), le Var, les Alpes-Maritimes puis la Haute-Savoie et le Rhône (majoritairement représenté par la ville de Lyon). Enfin, lorsque nous retirons les régions de Provence-Alpes-Côtes-d'Azur, l'Auvergne-Rhône-Alpes, l'Occitanie et la Corse, alors le paysage du reste de la France est très contrasté, tout n'est pas jaune, il y a de nouvelles minimales et maximales. **Les régions présentant un caractère frontalier** ont les loyers les plus élevés de France : près des frontières de l'Allemagne et de la Suisse, près de la région parisienne, proche des littoraux. Ainsi on peut plus facilement voir le département le moins cher de France selon nos données : **la Creuse**.
- Les infrastructures : de par la taille des cercles, nous voyons que le nombre le plus important d'infrastructures par habitant est plus élevé dans le sud de la France, soit les départements des Hautes-Alpes, les Pyrénées-Occidentales. Retirer le critère de l'Île-de-France ne fait pas changer cette première constatation. En revanche, enlever la Provence-Alpes-Côtes-d'Azur rend compte que les régions limitrophes à celle-ci sont bien pourvues en infrastructures. Le reste de la France est dans la moyenne basse. Pour continuer, en décochant les autres régions mentionnées, on constate que le **centre de la France offre plus d'infrastructures par habitant**, comme les départements de limitrophes à la Creuse l'Indre et la Corrèze. Cela s'explique certainement par le fait qu'il s'agisse d'une région moins densément peuplée. Enfin, les extrémités de la France, La Bretagne, le Grand-Est et le département des Pyrénées-Atlantiques sont également les mieux pourvus.
- Lecture croisée des loyers & infrastructures. Les zones régionales aux loyers très élevés proposent un nombre d'infrastructures assez réduit : l'Île-de-France en tête. Par comparaison, les zones départementales aux loyers moyennement élevés (mais tout de même au-dessus de la moyenne nationale) sont pourvues du plus grand nombre d'infrastructures loisirs : les Hautes-Alpes en tête. Enfin, les territoires aux loyers les moyennement élevés semblent être pourvus d'un nombre d'infrastructures moyens. Cet entre-deux ne semble pas exprimer beaucoup d'informations. C'est pourquoi, il faut continuer à sortir des filtres, les régions précédemment citées. On constate alors que les départements aux loyers les moins élevés sont pourvus du plus grand nombre d'infrastructures. Cette corrélation est à questionner d'un point de vue de l'utilisateur de City'nder. **Pourquoi vivre dans un département au loyer très élevé alors que l'offre sportive et culturelle est moins importante que dans les départements aux loyers très peu chers ? Est-ce seulement vrai que Paris est plus pauvre en offre culturelle et sportive que les départements du Centre-Val de Loire et de la Nouvelle-Aquitaine ? La réponse à cette question est certainement le résultat d'une faille de nos données alors synthétisé dans la conclusion.**



Carte : prix moyen des offres de loyer (maison et appartement) au m<sup>2</sup> et infrastructures "loisirs" par habitant sur l'ensemble du territoire métropolitain sans l'Île-de-France et la région Provence-Alpes-Côtes-d'Azur (capture d'écran centrée sur la France métropolitaine dont la Corse)



Carte : prix moyen des offres de loyer (maison et appartement) au m<sup>2</sup> et infrastructures "loisirs" par habitant sur l'ensemble du territoire métropolitain sans les régions Île-de-France, la Provence-Alpes-Côtes-d'Azur et Auvergne (capture d'écran centrée sur la France métropolitaine dont la Corse).

En guise de **conclusion**, nous constatons que le résultat obtenu est visuellement convaincant et facile d'utilisation. Cependant, il faut garder à l'esprit quelques points :

- Les biais :**
  - l'exactitude du résultat n'est pas garantie car l'exactitude des données ne l'est pas non plus ;
  - des données restrictives et non exhaustives car c'est le résultat d'un traitement excluant ;
  - un département de grande superficie et densément peuplé se voit favorisé par la division par population.
- D'autres possibilités** à explorer :
  - calculer le rapport d'infrastructure par population en fonction de la superficie par département. Cela permettrait de se rapprocher de la réalité du terrain. En effet, un territoire de grande superficie mais peu densément peuplé aura un nombre d'infrastructure/habitant plus élevé qu'un territoire extrêmement dense en population et concentré sur une petite superficie.
  - approfondir l'analyse des données par la définition des données extensives et intensives.

## 8. CONCLUSION

Pour conclure ce travail conséquent, nous souhaitons faire un rapport des atouts et inconvénients des outils explorés ainsi qu'un état des lieux de notre apprentissage sur la chaîne de traitement des données.

### 8.0.1. Utilisation des outils

- **Avantages**
  - **Un traitement de données conséquent facilité par Dataiku.** Nous nous sommes répartis de très grand set de données. *Dataiku* nous a permis de travailler séparément pour ensuite rassembler nos flux respectifs. L'outil permet une très bonne lisibilité du traitement. Nous avons donc pu aisément comprendre ce qui avait été fait par nos camarades. Nous pouvions également revenir en arrière à une étape précise du traitement. Dans son ensemble, *Dataiku* offre la possibilité de remonter à la source d'une longue chaîne de traitement, ce qui facilite la reprise des formules et recettes utilisées. Cela s'avère particulièrement pratique pour comprendre le cheminement des données et garantir la transparence du processus.
  - **Une visualisation des données aux nombreuses possibilités.** Nous avons pu réaliser des cartes assez facilement, ce qui est, par exemple, plus difficile avec des outils comme Excel. Nous avons également pu explorer les différents types de diagrammes rapidement en ne faisant que quelques manipulations.
  - *TableauPublic* a une grande capacité à **interroger les données à travers une variété de visualisations**. Bien que le travail avec des bases de données relationnelles implique souvent une concentration sur le traitement des données pour les rendre interrogables, c'est lors de la phase visuelle que certains problèmes émergent. *TableauPublic* nous offre la possibilité de mieux appréhender ces défis dès que nous avons une vision claire de notre objectif dans un jeu de données.
- **Inconvénients**
  - Dataiku est un outil qui fonctionne en **utilisant notre ordinateur comme serveur**. Or, nos datasets étaient très volumineux, ce qui a pu causer des problèmes parfois, pour ceux et celles d'entre nous dont les ordinateurs étaient peu puissants. Le traitement pouvait parfois être assez lent.
  - L'outil tableau est vraiment **difficile à prendre en main**. Il est facile de déplacer des données dans des champs pour essayer de visualiser quelque chose mais il y a un effet "trou noir", on ne comprend pas forcément ce que l'on fait - ou plutôt ce que Tableau réalise. Il y a une étape importante d'analyse à réaliser pour ne pas construire une visualisation inadaptée.
  - De plus, *TableauPublic* présente une **courbe d'apprentissage assez longue et exploratoire**. Bien que le logiciel propose une gamme variée de visualisations, son interface peut s'avérer moins intuitive, entraînant des difficultés et nécessitant un certain temps pour maîtriser les fonctionnalités de base. Ceci peut être moins attrayant pour les nouveaux utilisateurs. Pendant cette exploration, nous avons fait face à divers défis, en particulier liés à l'attribution automatique de *TableauPublic*. Les colonnes étaient en effet par défaut assignées en tant que sommes. Les échelles automatiques des graphiques présentaient également des difficultés, générant des incohérences entre les différentes visualisations. Enfin, certaines catégories que le logiciel propose sont entièrement tournées sur des thématiques anglo-saxonnes.
  - Ces obstacles ont apporté un éclairage nouveau sur l'analyse de données, révélant parfois des problèmes qui ne sont détectés qu'au moment de la visualisation. Par exemple,

**des superpositions des valeurs**, comme le nom de certaines villes homonymes telles que Beaulieu, peuvent entraver une analyse précise des données, impactant directement la représentation de la quantité d'hôtels ou de commerces, basée uniquement sur le nom de la commune.

#### 8.0.2. Apprentissages généraux sur la chaîne de traitement

- L'analyse de données exige **une rigueur à toute épreuve**, et ce tout le long de la chaîne de traitement.
- Ce travail nous a permis d'**appréhender le fonctionnement d'une chaîne de traitement de données**. Nous sommes partis de diverses tables contenant des données diverses pour les rassembler en une seule et être en mesure de les visualiser. Nous avons compris les dangers potentiels de ce type de traitement à grande échelle (biais dans l'analyse, manques de précision, ...)
- Malgré ces défis, la **data visualisation a suscité de nouvelles interrogations**, nous incitant souvent à réexaminer nos questions initiales à mesure que nos hypothèses se révélaient parfois incorrectes. Cette **expérience dynamique** a non seulement affiné notre compréhension des données analysées, mais a également souligné l'importance d'une **approche flexible dans le processus d'exploration**.
- Finalement, à travers cette démarche, nous avons **consolidé notre compréhension des nuances analytiques liées à nos données**, transcendant ainsi les simples chiffres pour révéler des *insights* pertinents. Ce travail collectif a renforcé notre maîtrise de la datavisualisation, ajoutant une dimension stratégique à notre apprentissage, au-delà de la simple représentation graphique.

Nous soutenons l'idée que ce travail a pour but de s'entraîner au processus du traitement et analyse des données plutôt que d'apporter un travail scientifique, rigoureux et vérifique s'approchant de la réalité.

#### 8.0.3. Problèmes et difficultés à recenser

- **Volume de données important au caractère muable.** La **diversité des données** que nous avons rassemblées fait la **force** mais aussi la **faiblesse** de notre traitement. En effet, une **quantité importante** de données signifie **davantage de probabilité d'imprécisions** dans le traitement.
- La principale **imprécision** à recenser concerne les **jointures**. Nous avons choisi les codes Insee comme clé de jointure car il s'agit de la colonne la plus fiable que nous avions dans le dataset de départ sur les loyers, la plus précise et la plus stable. Toutefois, chaque année, plus d'une dizaine de communes fusionnent ou sont supprimées car elles se regroupent en EPCI. Il peut donc y avoir des imprécisions concernant les communes qui ont récemment fusionné, ou des communes qui ont disparu depuis 2022. Lorsque les données concernent le présent, elles sont en constante évolution. Nous ne pouvions les capturer que sur un moment précis de l'histoire. Nous avons également conscience que, travaillant à l'échelle du territoire français dans sa globalité, nous ne pouvions être exhaustifs.

## 8.1. A RETENIR POUR L'AVENIR

Ce travail, éloigné de notre domaine d'expertise nous a finalement demandé un effort important de compréhension et d'apprentissage. Nous avons dû faire des recherches sur les communes françaises, le découpage territorial, les loyers, les statistiques voire les mathématiques, etc. Nous nous sommes tour à tour mis dans la peau de chefs de projets, de data scientists ou encore de concepteurs d'applications. Cette partie du projet constitue pour nous une belle invitation à poursuivre notre apprentissage, cette fois-ci appliqué à nos domaines d'expertise respectifs : l'histoire et l'histoire de l'art. Pour terminer, nous sommes convaincus que mettre en place cette chaîne de traitement est prometteur pour la prochaine étape de notre travail : le développement de l'application web City'nder.